Quantifying Synergistic Information

Thesis by

Virgil Griffith

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2013

(Submitted January 31, 2014)

This thesis would not exist were it not for many people. Some of the prominent ones are Douglas R. Hofstadter, for the inspiration; Giulio Tononi, for the theory; Christof Koch, for the ambition to explore new waters; and the DOE CSGF, for the funding to work in peace.

Acknowledgments

I wish to thank Christof Koch, Suzannah A. Fraker, Paul Williams, Mark Burgin, Tracey Ho, Edwin K. P. Chong, Christopher J. Ellison, Ryan G. James, Jim Beck, Shuki Bruck, and Pietro Perona.

Abstract

Within the microcosm of information theory, I explore what it means for a system to be functionally irreducible. This is operationalized as quantifying the extent to which cooperative or "synergistic" effects enable random variables $X_1, \ldots, X_n$ to predict (have mutual information about) a single target random variable Y. In Chapter 1, we introduce the problem with some emblematic examples. In Chapter 2, we show how six different measures from the existing literature fail to quantify this notion of synergistic mutual information. In Chapter 3, we take a step towards a measure of synergy which yields the first nontrivial lowerbound on synergistic mutual information. In Chapter 4, we find that synergy is but the weakest notion of a broader concept of *irreducibility*. In Chapter 5, we apply our results from Chapters 3 and 4 towards grounding Giulio Tononi's ambitious ϕ measure, which attempts to quantify the magnitude of consciousness experience.

Contents

Part I

Introducing the Problem

Chapter 1

What is Synergy?

The prior literature [24, 30, 1, 6, 19, 36] has termed several distinct concepts as "synergy". We define synergy as a special case of *irreducibility*—specifically, synergy is irreducibility to atomic elements. By definition, a group of two or more agents synergistically perform a task if and only if the performance of that task decreases when the agents work "separately", or in parallel isolation. It is important to remember that it is the collective *action* that is irreducible, not the agents themselves. A concrete example of irreducibility is the "agents" hydrogen and oxygen working to extinguish fire. Even when H_2 and O_2 are both present in the same container, if working separately neither extinguishes fire (on the contrary, fire grows!). But hydrogen and oxygen fused or "grouped" into a single entity, H_2O, readily extinguishes fire.

The concept of synergy spans many fields and theoretically could be applied to any non-subadditive function. But within the confines of Shannon information theory, synergy—or more formally, *synergistic information*—is a property of a set of n random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ cooperating to predict, that is, reduce the uncertainty of, a single target random variable Y.

1.1 Notation and PI-diagrams

We use the following notation throughout. Let

n: The number of predictors $X_1, X_2, \ldots, X_n$. $n \geq 2$.

$X_{1\ldots n}$: The joint random variable (cartesian product) of all n predictors $X_1 X_2 \ldots X_n$.

X_i: The i'th predictor random variable (r.v.). $1 \leq i \leq n$.

$\mathbf{X}$: The *set* of all n predictors $\{X_1, X_2, \ldots, X_n\}$.

Y: The *target r.v.* to be predicted.

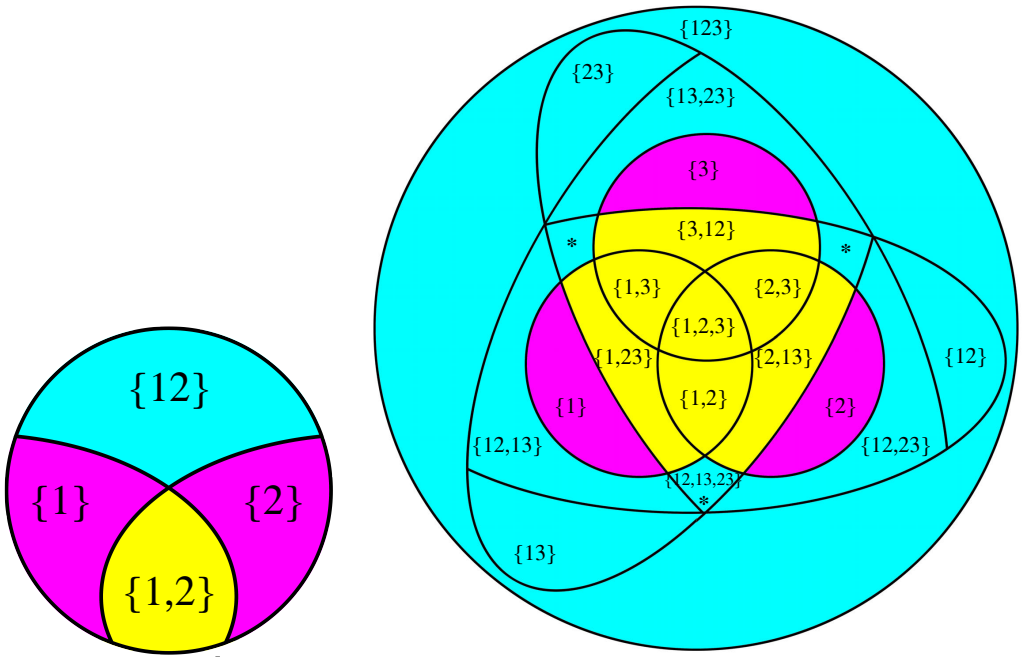
y: A particular state of the target r.v. Y.

All random variables are discrete, all logarithms are $\log_2$, and all calculations are in *bits*. Entropy and mutual information are as defined by [9], $\mathrm{H}(X) \equiv \sum_{x \in X} \Pr(x) \log \frac{1}{\Pr(x)}$, as well as $\mathrm{I}(X:Y) \equiv \sum_{x,y} \Pr(x,y) \log \frac{\Pr(x,y)}{\Pr(x)\Pr(y)}$.

1.1.1 Understanding PI-diagrams

Partial information diagrams (PI-diagrams), introduced by [36], extend Venn diagrams to properly represent synergy. Their framework has been invaluable to the evolution of our thinking on synergy.

A PI-diagram is composed of nonnegative *partial information regions* (PI-regions). Unlike the standard Venn entropy diagram in which the sum of all regions is the joint entropy $\mathrm{H}(X_{1\ldots n}, Y)$, in PI-diagrams the sum of all regions (i.e. the space of the PI-diagram) is the mutual information $\mathrm{I}(X_{1\ldots n}:Y)$. PI-diagrams are immensely helpful in understanding how the mutual information $\mathrm{I}(X_{1\ldots n}:Y)$ is distributed across the coalitions and singletons of $\mathbf{X}$.[1]



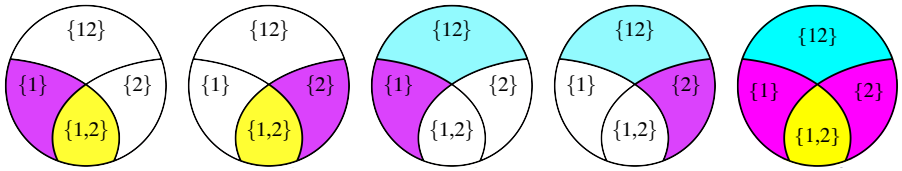
(a) $n = 2$ (b) $n = 3$

Figure 1.1: PI-diagrams for two and three predictors. Each PI-region represents nonnegative information about Y. A PI-region's color represents whether its information is redundant (yellow), unique (magenta), or synergistic (cyan). To preserve symmetry, the PI-region "$\{12, 13, 23\}$" is displayed as three separate regions each marked with a "*". All three *-regions should be treated as though they are a single region.

How to read PI-diagrams. Each PI-region is uniquely identified by its "set notation" where each element is denoted solely by the predictors' indices. For example, in the PI-diagram for $n = 2$

[1] Formally, how the mutual information is distributed across the set of all nonempty antichains on the powerset of $\mathbf{X}$[35].

(Figure 1.1a): $\{1\}$ is the information about Y only X_1 carries (likewise $\{2\}$ is the information only X_2 carries); $\{1,2\}$ is the information about Y that X_1 as well as X_2 carries, while $\{12\}$ is the information about Y that is specified only by the coalition (joint random variable) X_1X_2. For getting used to this way of thinking, common informational quantities are represented by colored regions in Figure 5.5.



(a) $\mathrm{I}(X_1:Y)$ (b) $\mathrm{I}(X_2:Y)$ (c) $\mathrm{I}(X_1:Y|X_2)$ (d) $\mathrm{I}(X_2:Y|X_1)$ (e) $\mathrm{I}(X_1X_2:Y)$

Figure 1.2: PI-diagrams for $n = 2$ representing standard informational quantities.

The general structure of a PI-diagram becomes clearer after examining the PI-diagram for $n = 3$ (Figure 1.1b). All PI-regions from $n = 2$ are again present. Each predictor (X_1, X_2, X_3) can carry unique information (regions labeled $\{1\}$, $\{2\}$, $\{3\}$), carry information redundantly with another predictor ($\{1,2\}$, $\{1,3\}$, $\{2,3\}$), or specify information through a coalition with another predictor ($\{12\}$, $\{13\}$, $\{23\}$). New in $n = 3$ is information carried by all three predictors ($\{1,2,3\}$) as well as information specified through a three-way coalition ($\{123\}$). Intriguingly, for three predictors, information can be provided by a coalition as well as a singleton ($\{1,23\}$, $\{2,13\}$, $\{3,12\}$) or specified by multiple coalitions ($\{12,13\}$, $\{12,23\}$, $\{13,23\}$, $\{12,13,23\}$).

1.2 Information can be redundant, unique, or synergistic

Each PI-region represents an irreducible nonnegative slice of the mutual information $\mathrm{I}(X_{1\ldots n}:Y)$ that is either:

1. **Redundant**. Information carried by a singleton predictor as well as available somewhere else. For $n = 2$: $\{1,2\}$. For $n = 3$: $\{1,2\}$, $\{1,3\}$, $\{2,3\}$, $\{1,2,3\}$, $\{1,23\}$, $\{2,13\}$, $\{3,12\}$.

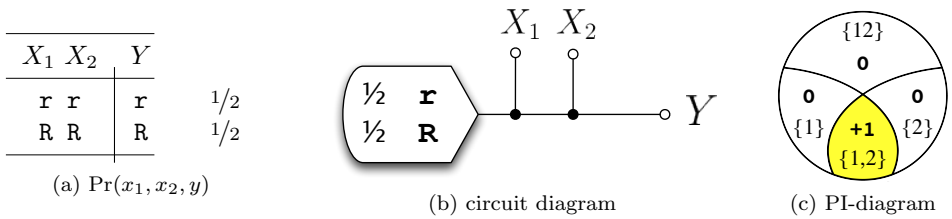
2. **Unique**. Information carried by exactly one singleton predictor and available nowhere else. For $n = 2$: $\{1\}$, $\{2\}$. For $n = 3$: $\{1\}$, $\{2\}$, $\{3\}$.

3. **Synergistic**. Any and all information in $\mathrm{I}(X_{1\ldots n}:Y)$ that is not carried by a singleton predictor. $n = 2$: $\{12\}$. For $n = 3$: $\{12\}$, $\{13\}$, $\{23\}$, $\{123\}$, $\{12,13\}$, $\{12,23\}$, $\{13,23\}$, $\{12,13,23\}$.

Although a single PI-region is either redundant, unique, or synergistic, a single state of the target can have any combination of positive PI-regions, i.e. a single state of the target can convey redundant, unique, and synergistic information. This surprising fact is demonstrated in Figure 3.4.

1.2.1 Example Rdn: Redundant information

If X_1 and X_2 carry some of the same information[2] (reduce the same uncertainty) about Y, then we say the set $\mathbf{X} = \{X_1, X_2\}$ has some *redundant information* about Y. Figure 1.3 illustrates a simple case of redundant information. Y has two equiprobable states: `r` and `R` (`r`/`R` for "redundant bit"). Examining X_1 or X_2 identically specifies one bit of Y, thus we say set $\mathbf{X} = \{X_1, X_2\}$ has one bit of redundant information about Y.



(a) $\Pr(x_1, x_2, y)$ (b) circuit diagram (c) PI-diagram

Figure 1.3: Example RDN. Figure 1.3a shows the joint distribution of r.v.'s X_1, X_2, and Y, and the joint probability $\Pr(x_1, x_2, y)$ is along the right-hand side of (a), revealing that all three terms are fully correlated. Figure 1.3b represents the joint distribution as an electrical circuit. Figure 1.3c is the PI-diagram indicating that set $\{X_1, X_2\}$ has 1 bit of redundant information about Y. $\mathrm{I}(X_1X_2{:}Y) = \mathrm{I}(X_1{:}Y) = \mathrm{I}(X_2{:}Y) = \mathrm{H}(Y) = 1$ bit.

1.2.2 Example Unq: Unique information

Predictor X_i carries *unique information* about Y if and only if X_i specifies information about Y that is not specified by anything else (a singleton or coalition of the other $n-1$ predictors). Figure 1.4 illustrates a simple case of unique information. Y has four equiprobable states: `ab`, `aB`, `Ab`, and `AB`. X_1 uniquely specifies bit `a`/`A`, and X_2 uniquely specifies bit `b`/`B`. If we had instead labeled the Y-states: `0`, `1`, `2`, and `3`, X_1 and X_2 would still have strictly unique information about Y. The state of X_1 would specify between $\{0, 1\}$ and $\{2, 3\}$, and the state of X_2 would specify between $\{0, 2\}$ and $\{1, 3\}$—together fully specifying the state of Y.

1.2.3 Example Xor: Synergistic information

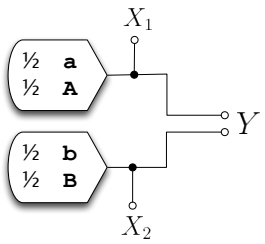
A set of predictors $\mathbf{X} = \{X_1, \ldots, X_n\}$ has synergistic information about Y if and only if the whole $(X_{1\ldots n})$ specifies information about Y that is not specified by any singleton predictor. The canonical example of synergistic information is the XOR-gate (Figure 1.5). In this example, the whole X_1X_2 fully specifies Y,

$$\mathrm{I}(X_1X_2{:}Y) = \mathrm{H}(Y) = 1 \text{ bit},$$

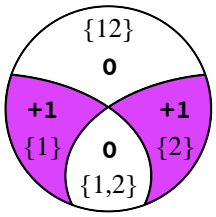
[2] X_1 and X_2 providing identical information about Y is different from providing the same *magnitude* of information about Y, i.e. $\mathrm{I}(X_1{:}Y) = \mathrm{I}(X_2{:}Y)$. Example UNQ (Figure 1.4) is an example where $\mathrm{I}(X_1{:}Y) = \mathrm{I}(X_2{:}Y) = 1$ bit yet X_1 and X_2 specify "different bits" of Y. Providing the same magnitude of information about Y is neither necessary or sufficient for providing some identical information about Y.

X_1	X_2	Y	
a	b	ab	1/4
a	B	aB	1/4
A	b	Ab	1/4
A	B	AB	1/4

(a) $\Pr(x_1, x_2, y)$



(b) circuit diagram



(c) PI-diagram

Figure 1.4: Example UNQ. X_1 and X_2 each uniquely specify a single bit of Y. $\mathrm{I}(X_1X_2{:}Y) = \mathrm{H}(Y) = 2$ bits. The joint probability $\Pr(x_1, x_2, y)$ is along the right-hand side of (a).

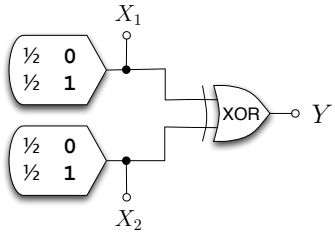
but the singletons X_1 and X_2 specify *nothing* about Y,

$$\mathrm{I}(X_1{:}Y) = \mathrm{I}(X_2{:}Y) = 0 \text{ bits.}$$

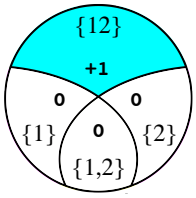
With both X_1 and X_2 themselves having zero information about Y, we know that there can not be any redundant or unique information about Y, that the three PI-regions $\{1\} = \{2\} = \{1,2\} = 0$ bits. As the information between X_1X_2 and Y must come from somewhere, by elimination we conclude that X_1 and X_2 synergistically specify Y.

X_1	X_2	Y	
0	0	0	1/4
0	1	1	1/4
1	0	1	1/4
1	1	0	1/4

(a) $\Pr(x_1, x_2, y)$



(b) circuit diagram



(c) PI-diagram

Figure 1.5: Example XOR. X_1 and X_2 synergistically specify Y. $\mathrm{I}(X_1X_2{:}Y) = \mathrm{H}(Y) = 1$ bit. The joint probability $\Pr(x_1, x_2, y)$ is along the right-hand side of (a).

Chapter 2

Six Prior Measures of Synergy

2.1 Definitions

2.1.1 Multivariate Mutual Information: $\mathrm{MMI}(X_1;\cdots;X_n;Y)$

The first information-theoretic measure of synergy dates to 1954 from [24]. Inspired by Venn entropy diagrams, they defined the *multivariate mutual information* (MMI), $\mathrm{MMI}(X_1;\cdots;X_n;Y) \equiv \sum_{T\subseteq\{X_1,\ldots,X_n,Y\}}(-1)^{|T|+1}\mathrm{H}(T)$. Negative MMI was understood to be synergy. Therefore the MMI measure of synergy is,

$$\begin{aligned}\mathcal{S}_{\mathrm{MMI}}(X_1;\cdots;X_n;Y) &\equiv -\sum_{\mathbf{T}\subseteq\{X_1,\ldots,X_n,Y\}}(-1)^{|\mathbf{T}|+1}\mathrm{H}(T)\\ &= \sum_{\mathbf{T}\subseteq\{X_1,\ldots,X_n,Y\}}(-1)^{|\mathbf{T}|}\mathrm{H}(T)\ .\end{aligned} \tag{2.1}$$

2.1.2 Interaction Information: $\mathcal{II}\,(X_1;\cdots;X_n;Y)$

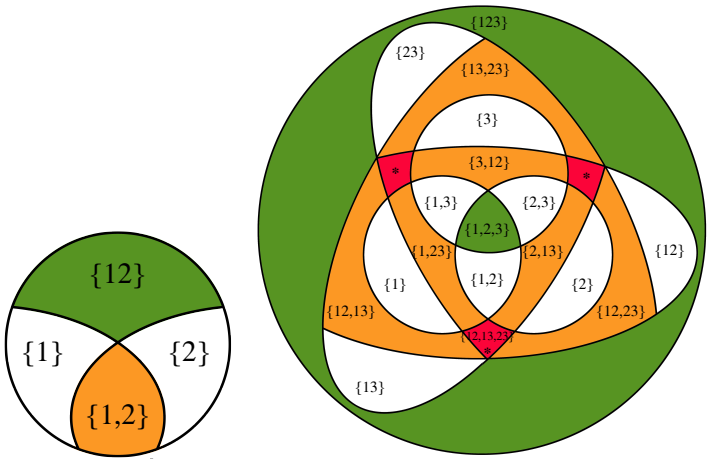
Interaction information ($\mathcal{II}$), sometimes called the co-information, was introduced in [6] and tweaks MMI synergy measure. Although intended to measure informational "groupness" [6], Interaction Information is commonly interpreted as the magnitude of "information bound up in a set of variables, beyond that which is present in any subset of those variables."[1]

Interaction Information among the n predictors and Y is defined as,

$$\begin{aligned}\mathcal{II}\,(X_1;\cdots;X_n;Y) &\equiv (-1)^n\,\mathcal{S}_{\mathrm{MMI}}(X_1;\cdots;X_n;Y)\\ &= \sum_{\mathbf{T}\subseteq\{X_1,\ldots,X_n,Y\}}(-1)^{n-|\mathbf{T}|}\mathrm{H}(T)\ .\end{aligned} \tag{2.2}$$

[1] From `http://en.wikipedia.org/wiki/Interaction_information`.

Interaction Information is a signed measure where a positive value signifies synergy and a negative value signifies redundancy. Representing Interaction Information as a PI-diagram (Figure 2.1) reveals an intimidating imbroglio of added and subtracted PI-regions.



(a) $\mathcal{II}(\{X_1, X_2\} : Y)$ (b) $\mathcal{II}(X_1; X_2; X_3; Y)$

Figure 2.1: PI-diagrams illustrating interaction information for $n = 2$ (left) and $n = 3$ (right). The colors denote the added and subtracted PI-regions. $\mathrm{WMS}(\mathbf{X} : Y)$ is the green PI-region(s), minus the orange PI-region(s), minus two times any red PI-region.

2.1.3 WholeMinusSum synergy: $\mathrm{WMS}(\mathbf{X} : Y)$

The earliest known sightings of bivariate WholeMinusSum synergy (WMS) are in [13, 12], with the general case in [11]. WholeMinusSum synergy is a signed measure where a positive value signifies synergy and a negative value signifies redundancy. WholeMinusSum synergy is defined by eq. (2.3) and interestingly reduces to eq. (2.5)—the difference of two *total correlations*.[2]

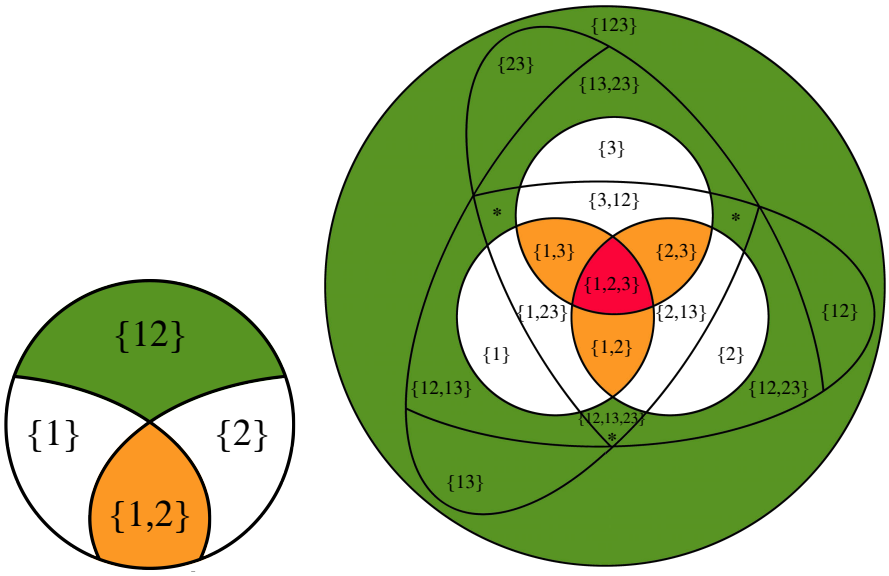
$$\begin{aligned}
\mathrm{WMS}(\mathbf{X} : Y) &\equiv \mathrm{I}(X_{1\ldots n} : Y) - \sum_{i=1}^{n} \mathrm{I}(X_i : Y) && (2.3)\\
&= \sum_{i=1}^{n} \mathrm{H}\big(X_i|Y\big) - \mathrm{H}\big(X_{1\ldots n}|Y\big) - \left[\sum_{i=1}^{n} \mathrm{H}(X_i) - \mathrm{H}(X_{1\ldots n})\right] && (2.4)\\
&= \mathrm{TC}(X_1; \cdots; X_n|Y) - \mathrm{TC}(X_1; \cdots; X_n) && (2.5)
\end{aligned}$$

Representing eq. (2.3) for $n = 2$ as a PI-diagram (Figure 2.2a) reveals that WMS is the synergy between X_1 and X_2 *minus* their redundancy. Thus, when there is an equal magnitude of synergy and redundancy between X_1 and X_2, WholeMinusSum synergy is *zero*—leading one to *erroneously*

[2] $\mathrm{TC}(X_1; \cdots; X_n) = -\mathrm{H}(X_{1\ldots n}) + \sum_{i=1}^{n} \mathrm{H}(X_i)$ per [17].

conclude there is no synergy or redundancy present.[3]

The PI-diagram for $n = 3$ (Figure 2.2b) reaveals that WholeMinusSum double-subtracts PI-regions {1,2}, {1,3}, {2,3} and triple-subtracts PI-region {1,2,3}, revealing that for $n > 2$ WMS $(\mathbf{X} : Y)$ becomes synergy minus the redundancy *counted multiple times*.



(a) WMS $(\{X_1, X_2\} : Y)$ (b) WMS $(\{X_1, X_2, X_3\} : Y)$

Figure 2.2: PI-diagrams illustrating WholeMinusSum synergy for $n = 2$ (left) and $n = 3$ (right). The colors denote the added and subtracted PI-regions. WMS $(\mathbf{X} : Y)$ is the green PI-region(s) minus the orange PI-region(s) minus two times any red PI-region.

2.1.4 WholeMinusPartitionSum: WMPS $(\mathbf{X} : Y)$

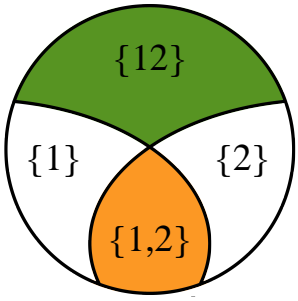
WholeMinusPartitionSum, denoted WMPS $(\mathbf{X} : Y)$, is a stricter generalization of WMS synergy for $n > 2$. It was introduced in [34, 1] and is defined as,

$$\text{WMPS}\,(\mathbf{X} : Y) \equiv \text{I}(X_{1\ldots n} : Y) - \max_{\mathbf{P}} \sum_{i=1}^{|\mathbf{P}|} \text{I}(P_i : Y) \ , \tag{2.6}$$

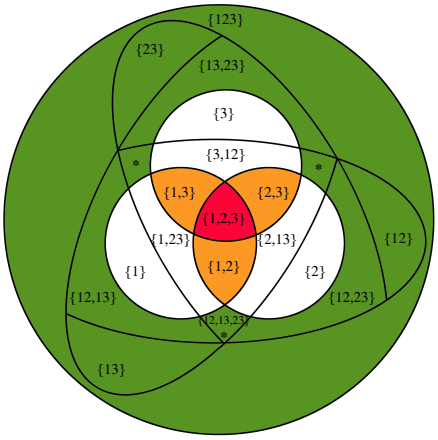
where $\mathbf{P}$ enumerates over all partitions of the set of predictors $\{X_1, \ldots, X_n\}$. WholeMinusPartitionSum is a signed measure where a positive value signifies synergy and a negative value signifies redundancy.

For $n = 3$, there are four partitions of $\mathbf{X}$ resulting in four possible PI-diagrams—one for each partition. Figure 2.3 depicts the four possible values of WMPS$(\{X_1, X_2, X_3\} : Y)$. Because $\{X_1, \ldots, X_n\}$ is a possible partition of $\mathbf{X}$, WMPS$(\mathbf{X} : Y) \leq$ WMS$(\mathbf{X} : Y)$.

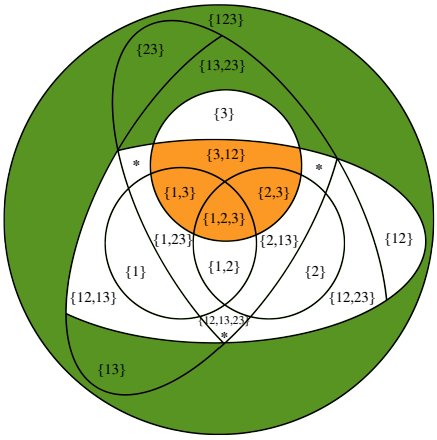
[3]This is deeper than [29]'s point that a mish-mash of synergy and redundancy across different states of $y \in Y$ can average to zero. E.g., Figure 2.6 evaluates to zero for *every state* $y \in Y$.



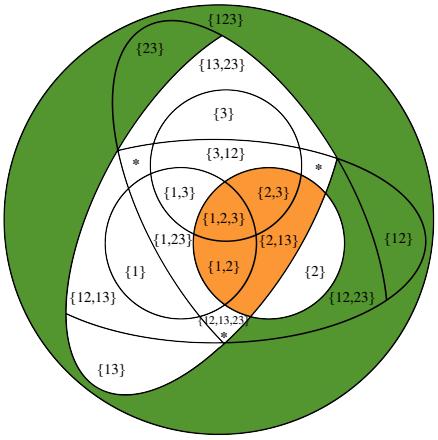
(a) WMPS $\left(\{X_1, X_2\} : Y\right)$



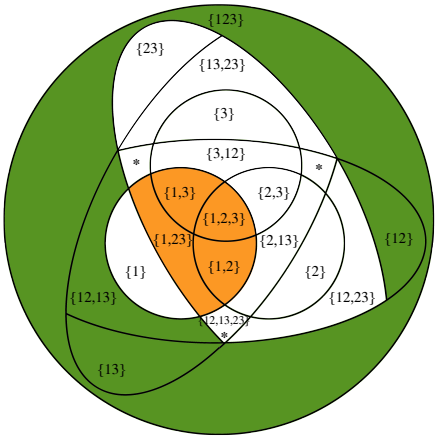
(b) $\mathbf{P} = \{X_1, X_2, X_3\}$



(c) $\mathbf{P} = \{X_1X_2, X_3\}$



(d) $\mathbf{P} = \{X_1X_3, X_2\}$



(e) $\mathbf{P} = \{X_2X_3, X_1\}$

Figure 2.3: PI-diagrams depicting WholeMinusPartitionSum synergy for $n = 2$ (2.3a) and $n = 3$ (2.3b–2.3e). Each measure is the green PI-regions minus the orange PI-regions minus two times any red PI-region. WMPS $\left(\{X_1, X_2, X_3\} : Y\right)$ is the *minimum* value over subfigures 2.3b–2.3e.

2.1.5 $\mathrm{I}_{\max}$ synergy: $\mathcal{S}_{\max}\left(\mathbf{X} : Y\right)$

$\mathrm{I}_{\max}$ synergy, denoted $\mathcal{S}_{\max}$, was the first synergy measure derived from Partial Information Decomposition[36]. $\mathcal{S}_{\max}$ defines synergy as the whole beyond the state-dependent *maximum* of its elements,

$$\mathcal{S}_{\max}(\mathbf{X}:Y) \equiv \mathrm{I}(X_{1\ldots n}:Y) - \mathrm{I}_{\max}\left(\{X_1,\ldots,X_n\}:Y\right) \tag{2.7}$$

$$= \mathrm{I}(X_{1\ldots n}:Y) - \sum_{y\in Y}\Pr(Y=y)\max_i \mathrm{I}(X_i:Y=y)\ , \tag{2.8}$$

where $\mathrm{I}(X_i:Y=y)$ is [10]'s "specific-surprise",

$$\mathrm{I}(X_i:Y=y) \equiv \mathrm{D}_{\mathrm{KL}}\left[\Pr\big(X_i|y\big)\middle\|\Pr(X_i)\right] \tag{2.9}$$

$$= \sum_{x_i\in X_i}\Pr\big(x_i|y\big)\log\frac{\Pr(x_i,y)}{\Pr(x_i)\Pr(y)}\ . \tag{2.10}$$

There are two major advantages of $\mathcal{S}_{\max}$ synergy. $\mathcal{S}_{\max}$ is not only nonnegative, but also invariant to duplicate predictors.

2.1.6 Correlational importance: $\Delta\,\mathrm{I}\,(\mathbf{X};Y)$

Correlational importance, denoted $\Delta\,\mathrm{I}$, comes from [27, 25, 26, 28, 21]. Correlational importance quantifies the "informational importance of conditional dependence" or the "information lost when ignoring conditional dependence" among the predictors decoding target Y. On casual inspection, $\Delta\,\mathrm{I}$ seems related to our intuitive conception of synergy. $\Delta\,\mathrm{I}$ is defined as,

$$\Delta\,\mathrm{I}\,(\mathbf{X};Y) \equiv \mathrm{D}_{\mathrm{KL}}\left[\Pr\big(Y|X_{1\ldots n}\big)\middle\|\Pr_{\mathrm{ind}}(Y|\mathbf{X})\right] \tag{2.11}$$

$$= \sum_{y,\mathbf{x}\in Y,\mathbf{X}}\Pr(y,x_{1\ldots n})\log\frac{\Pr\big(y|x_{1\ldots n}\big)}{\Pr_{\mathrm{ind}}(y|\mathbf{x})}\ , \tag{2.12}$$

where $\Pr_{\mathrm{ind}}\big(y|\mathbf{x}\big) \equiv \frac{\Pr(y)\prod_{i=1}^n \Pr\big(x_i|y\big)}{\sum_{y'}\Pr(y')\prod_{i=1}^n \Pr\big(x_i|y'\big)}$. After some algebra[4] eq. (2.12) becomes,

$$\Delta\,\mathrm{I}\,(\mathbf{X};Y) = \mathrm{TC}\,(X_1;\cdots;X_n|Y) - \mathrm{D}_{\mathrm{KL}}\left[\Pr(X_{1\ldots n})\middle\|\sum_y \Pr(y)\prod_{i=1}^n \Pr\big(X_i|y\big)\right]\ . \tag{2.13}$$

$\Delta\,\mathrm{I}$ is conceptually innovative, yet examples reveal that $\Delta\,\mathrm{I}$ measures something ever-so-subtly different from intuitive synergistic information.

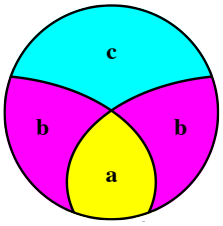
[4] See Appendix 2.A for the steps between eqs. (2.12) and (2.13).

2.2 The six prior measures are not equivalent

For $n = 2$, the four measures $\mathcal{S}_{\text{MMI}}$, $\mathcal{II}$, WMS, and WMPS are equivalent. But in general, none of these six measures are equivalent. Example AND (Figure 2.4) shows that $\mathcal{S}_{\max}$ and $\Delta\,\text{I}$ are not equivalent. Example XORMULTICOAL (Figure 2.5) shows that $\mathcal{S}_{\text{MMI}}$, $\mathcal{II}$, WMS, and WMPS are not equivalent.

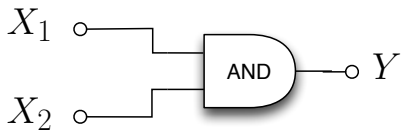
X_1	X_2	Y	
0	0	0	1/4
0	1	0	1/4
1	0	0	1/4
1	1	1	1/4

(a) $\Pr(x_1, x_2, y)$



(b) PI-diagram

$$0.189 \leq c \leq 0.5$$
$$0 \leq b \leq 0.311$$
$$0 \leq a \leq 0.311$$



(c) circuit diagram

Figure 2.4: Example AND. The exact PI-decomposition of an AND-gate remains uncertain. But we can bound a, b, and c using WMS and $\mathcal{S}_{\max}$.

Example	$\mathcal{S}_{\text{MMI}}$	$\mathcal{II}$	WMS	WMPS	$\mathcal{S}_{\max}$	$\Delta\,\text{I}$
AND	0.189	0.189	0.189	0.189	1/2	0.104
XORMULTICOAL	2	–2	1	0	1	1

Table 2.1: Examples demonstrating that the six prior measures are not equivalent.

2.3 Counter-intuitive behaviors of the six prior measures

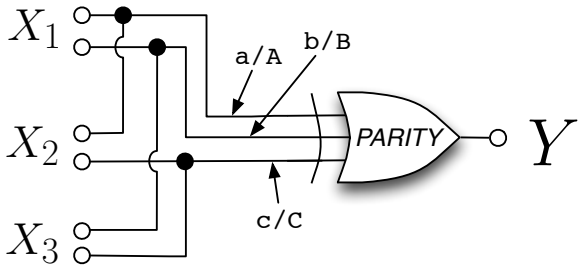
2.3.1 $\text{I}_{\max}$ synergy: $\mathcal{S}_{\max}$

Despite several desired properties, $\mathcal{S}_{\max}$ sometimes miscategorizes merely unique information as synergistic. This can be seen in example UNQ (Figure 1.4). In example UNQ, the wires in Figure 1.4b don't even touch, yet $\mathcal{S}_{\max}$ asserts there is one bit of synergy and one bit of redundancy—this is palpably strange.

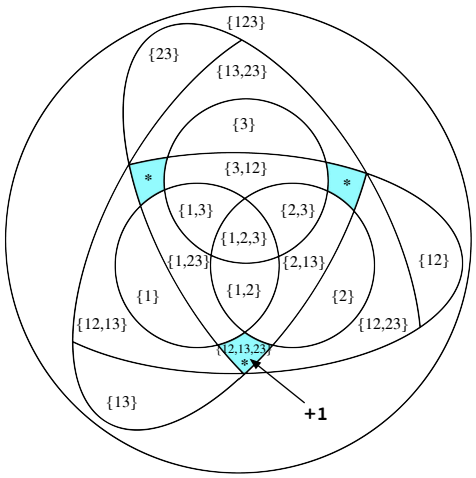
A more abstract way to understand why $\mathcal{S}_{\max}$ overestimates synergy is to imagine a hypothetical example where there are exactly two bits of unique information for every state $y \in Y$ and no synergy or redundancy. $\mathcal{S}_{\max}$ would be the whole (both unique bits) minus the *maximum* over both

X_1 X_2 X_3	Y	
ab ac bc	0	1/8
AB Ac Bc	0	1/8
Ab AC bC	0	1/8
aB aC BC	0	1/8
Ab Ac bc	1	1/8
aB ac Bc	1	1/8
ab aC bC	1	1/8
AB AC BC	1	1/8

(a) $\Pr(x_1, x_2, x_3, y)$



(b) circuit diagram



(c) PI-diagram

Figure 2.5: Example XorMultiCoal demonstrates how the same information can be specified by multiple coalitions. In XorMultiCoal the target Y has one bit of uncertainty, $\mathrm{H}(Y) = 1$ bit, and Y is the *parity* of three incoming wires. Just as the output of Xor is specified only after knowing the state of both inputs, the output of XorMultiCoal is specified only after knowing the state of all three wires. Each predictor is distinct and has access to two of the three incoming wires. For example, predictor X_1 has access to the `a/A` and `b/B` wires, X_2 has access to the `a/A` and `c/C` wires, and X_3 has access to the `b/B` and `c/C` wires. Although no single predictor specifies Y, any coalition of two predictors has access to all three wires and fully specifies Y, $\mathrm{I}(X_1X_2{:}Y) = \mathrm{I}(X_1X_3{:}Y) = \mathrm{I}(X_2X_3{:}Y) = \mathrm{H}(Y) = 1$ bit. In the PI-diagram this puts one bit in PI-region $\{12, 13, 23\}$ and zero everywhere else.

predictors, which would be the $\max[1, 1] = 1$ bit. The $S_{\max}$ synergy would then be $2 - 1 = 1$ bit of synergy, even though by definition there was no synergy, but merely two bits of unique information.

Altogether, we conclude that $\mathcal{S}_{\max}$ *overestimates* the intuitive synergy by miscategorizing merely unique information as synergistic whenever two or more predictors have unique information about the target.

2.3.2 $\mathcal{S}_{\mathrm{MMI}}$, $\mathcal{II}$, WMS, WMPS

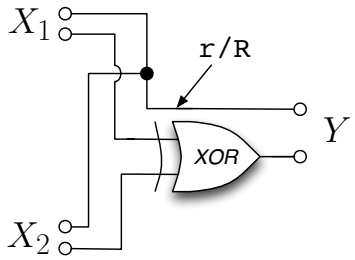
All four of these measures are equivalent for $n = 2$. Given this agreement, it is ironic that there are counter-intuitive examples even for $n = 2$. A concrete example demonstrating a "synergy minus

redundancy" behavior for $n = 2$ is example RdnXor (Figure 2.6), which overlays examples Rdn and Xor to form a single system. The target Y has two bits of uncertainty, i.e. $\mathrm{H}(Y) = 2$. Like Rdn, either X_1 or X_2 identically specifies the letter of Y (r/R), making one bit of redundant information. Like Xor, only the coalition X_1X_2 specifies the digit of Y (0/1), making one bit of synergistic information. Together this makes one bit of redundancy and one bit of synergy. We assert that for $n = 2$, all four measures *underestimate* the synergy. Equivalently, we say that their answer for $n = 2$ is a *lowerbound* on the intuitive synergy.

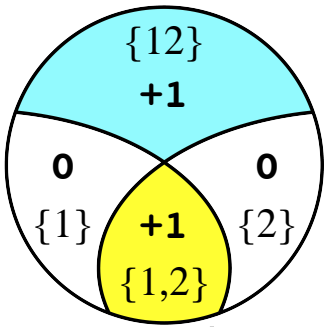
Note that in RdnXor every state $y \in Y$ conveys one bit of redundant information and one bit of synergistic information, e.g. for the state $y =$ r0 the letter "r" is specified redundantly and the digit "0" is specified synergistically.

X_1	X_2	Y	
r0	r0	r0	1/8
r0	r1	r1	1/8
r1	r0	r1	1/8
r1	r1	r0	1/8
R0	R0	R0	1/8
R0	R1	R1	1/8
R1	R0	R1	1/8
R1	R1	R0	1/8

(a) $\Pr(x_1, x_2, y)$



(b) circuit diagram



(c) PI-diagram

Figure 2.6: Example RdnXor has one bit of redundancy and one bit of synergy. Yet for this example, the four most common measures of synergy arrive at zero bits.

Our next example, PartyRdnRdn (Figure 2.7), has one bit of synergy and two bits of redundancy for a total of $\mathrm{I}(X_1X_2X_3 : Y) = \mathrm{H}(Y) = 3$ bits. It emphasizes the disagreement between $\mathcal{II}$ and measures $\mathcal{S}_{\mathrm{MMI}}$, WMS, and WMPS. If $\mathcal{S}_{\mathrm{MMI}}$, WMS, or WMPS were always simply "synergy minus redundancy", then one of them would calculate $1 - 2 = -1$ bits. But for this example all three measures subtracts redundancies *multiple times* to calculate $1 - (2 \cdot 2) = -3$ bits, signifying all three bits of $\mathrm{H}(Y)$ are specified redundantly. $\mathcal{II}$ makes a different misstep. Instead of subtracting redundancy multiple times, for $n = 3$ $\mathcal{II}$ *adds* the maximum redundancy to calculate $1 + 2 = +3$ bits, signifying three bits of synergy and no redundancy. Both answers are palpably mistaken.

2.3.3 Correlational importance: $\Delta\,\mathrm{I}$

The first concerning example is [29]'s Figure 4, where $\Delta\,\mathrm{I}$ exceeds the mutual information $\mathrm{I}(X_{1\ldots n} : Y)$ with $\Delta\,\mathrm{I}(\mathbf{X}; Y) = 0.0145$ and $\mathrm{I}(X_{1\ldots n} : Y) = 0.0140$. This fact alone prevents interpreting $\Delta\,\mathrm{I}$ the magnitude of mutual $\mathrm{I}(X_{1\ldots n} : Y)$ arising from correlational dependence.

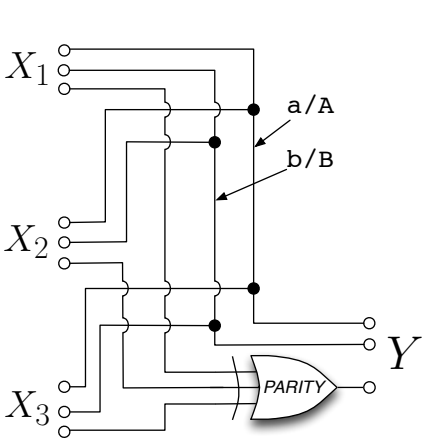
Could $\Delta\,\mathrm{I}$ upperbound synergy instead? We turn to example And (Figure 2.4) with $n = 2$ independent binary predictors and target Y is the AND of X_1 and X_2. Although And's exact

PI-region decomposition remains uncertain, we can still bound the synergy. For example, the $\mathrm{WMS}(\{X_1, X_2\} : Y) \approx 0.189$ and $\mathcal{S}_{\max}\left(\{X_1, X_2\} : Y\right) = 0.5$ bits. So we know the synergy must be between $(0.189, 0.5]$ bits. Despite this, $\Delta\,\mathrm{I}\,(\mathbf{X}; Y) = 0.104$ bits, thus $\Delta\,\mathrm{I}$ does not upperbound synergy either.

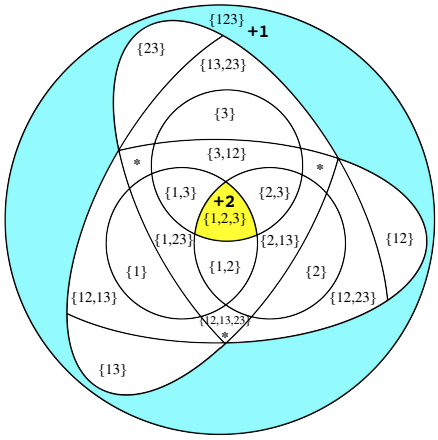
Taking both together, we conclude that $\Delta\,\mathrm{I}$ measures something fundamentally different from synergistic information.

Example	$\mathcal{S}_{\mathrm{MMI}}$	$\mathcal{II}$	WMS	WMPS	$\mathcal{S}_{\max}$	$\Delta\,\mathrm{I}$
Unq	0	0	0	0	1	0
RdnXor	0	0	0	0	1	1
ParityRdnRdn	–3	3	–3	–3	1	1
And	0.189	0.189	0.189	0.189	1/2	0.104

Table 2.2: Examples demonstrating that all six prior measures have shortcomings.



(a) circuit diagram



(c) PI-diagram

X_1	X_2	X_3	Y	
ab0	ab0	ab0	ab0	1/32
ab0	ab1	ab1	ab0	1/32
ab1	ab0	ab1	ab0	1/32
ab1	ab1	ab0	ab0	1/32
aB0	aB0	aB0	aB0	1/32
aB0	aB1	aB1	aB0	1/32
aB1	aB0	aB1	aB0	1/32
aB1	aB1	aB0	aB0	1/32
ab0	ab0	ab1	ab1	1/32
ab0	ab1	ab0	ab1	1/32
ab1	ab0	ab0	ab1	1/32
ab1	ab1	ab1	ab1	1/32
aB0	aB0	aB1	aB1	1/32
aB0	aB1	aB0	aB1	1/32
aB1	aB0	aB0	aB1	1/32
aB1	aB1	aB1	aB1	1/32

X_1	X_2	X_3	Y	
Ab0	Ab0	Ab0	Ab0	1/32
Ab0	Ab1	Ab1	Ab0	1/32
Ab1	Ab0	Ab1	Ab0	1/32
Ab1	Ab1	Ab0	Ab0	1/32
AB0	AB0	AB0	AB0	1/32
AB0	AB1	AB1	AB0	1/32
AB1	AB0	AB1	AB0	1/32
AB1	AB1	AB0	AB0	1/32
Ab0	Ab0	Ab1	Ab1	1/32
Ab0	Ab1	Ab0	Ab1	1/32
Ab1	Ab0	Ab0	Ab1	1/32
Ab1	Ab1	Ab1	Ab1	1/32
AB0	AB0	AB1	AB1	1/32
AB0	AB1	AB0	AB1	1/32
AB1	AB0	AB0	AB1	1/32
AB1	AB1	AB1	AB1	1/32

(b) $\Pr(x_1, x_2, x_3, y)$

Figure 2.7: Example PARITYRDNRDN. Three predictors redundantly specify two bits of Y, $\mathrm{I}(X_1 : Y) = \mathrm{I}(X_2 : Y) = \mathrm{I}(X_3 : Y) = 2$ bits. At the same time, the three predictors holistically specify the third and final bit of Y, $\mathrm{I}(X_1 X_2 X_3 : Y) = \mathrm{H}(Y) = 3$ bits.

Appendix

2.A Algebraic simplification of ΔI

Prior literature [25, 26, 28, 21] defines $\Delta\,\mathrm{I}\,(\mathbf{X};Y)$ as,

$$\Delta\,\mathrm{I}\,(\mathbf{X};Y) \equiv \mathrm{D_{KL}}\left[\Pr\big(Y|X_{1\ldots n}\big)\middle\|\Pr_{\mathrm{ind}}\,(Y|\mathbf{X})\right] \tag{2.14}$$

$$= \sum_{\mathbf{x},y\in\mathbf{X},Y} \Pr(\mathbf{x},y)\log\frac{\Pr\big(y|\mathbf{x}\big)}{\Pr_{\mathrm{ind}}(y|\mathbf{x})}\,. \tag{2.15}$$

Where,

$$\Pr_{\mathrm{ind}}(Y=y|\mathbf{X}=\mathbf{x}) \equiv \frac{\Pr(y)\Pr_{\mathrm{ind}}(\mathbf{X}=\mathbf{x}|Y=y)}{\Pr_{\mathrm{ind}}(\mathbf{X}=\mathbf{x})} \tag{2.16}$$

$$= \frac{\Pr(y)\prod_{i=1}^{n}\Pr\big(x_i|y\big)}{\Pr_{\mathrm{ind}}(\mathbf{x})} \tag{2.17}$$

$$\Pr_{\mathrm{ind}}(\mathbf{X}=\mathbf{x}) \equiv \sum_{y\in Y}\Pr(Y=y)\prod_{i=1}^{n}\Pr\big(x_i|y\big) \tag{2.18}$$

The definition of $\Delta\,\mathrm{I}$, eq. (2.14), reduces to,

$$
\begin{aligned}
\Delta\,\mathrm{I}\left(\mathbf{X};Y\right) &= \sum_{\mathbf{x},y\in\mathbf{X},Y}\Pr(\mathbf{x},y)\log\frac{\Pr\big(y|\mathbf{x}\big)}{\Pr_{\mathrm{ind}}(y|\mathbf{x})} && (2.19)\\
&= \sum_{\mathbf{x},y\in\mathbf{X},Y}\Pr(\mathbf{x},y)\log\frac{\Pr\big(y|\mathbf{x}\big)\Pr_{\mathrm{ind}}(\mathbf{x})}{\Pr(y)\prod_{i=1}^{n}\Pr\big(x_i|y\big)} && (2.20)\\
&= \sum_{\mathbf{x},y\in\mathbf{X},Y}\Pr(\mathbf{x},y)\log\frac{\Pr\big(\mathbf{x}|y\big)}{\prod_{i=1}^{n}\Pr\big(x_i|y\big)}\frac{\Pr_{\mathrm{ind}}(\mathbf{x})}{\Pr(\mathbf{x})} && (2.21)\\
&= \sum_{\mathbf{x},y\in\mathbf{X},Y}\Pr(\mathbf{x},y)\log\frac{\Pr\big(\mathbf{x}|y\big)}{\prod_{i=1}^{n}\Pr\big(x_i|y\big)} + \sum_{\mathbf{x},y\in\mathbf{X},Y}\Pr(\mathbf{x},y)\log\frac{\Pr_{\mathrm{ind}}(\mathbf{x})}{\Pr(\mathbf{x})}\\
&= \sum_{\mathbf{x},y\in\mathbf{X},Y}\Pr(\mathbf{x},y)\log\frac{\Pr\big(\mathbf{x}|y\big)}{\prod_{i=1}^{n}\Pr\big(x_i|y\big)} - \sum_{\mathbf{x}\in\mathbf{X}}\Pr(\mathbf{x})\log\frac{\Pr(\mathbf{x})}{\Pr_{\mathrm{ind}}(\mathbf{x})} && (2.22)\\
&= \mathrm{D}_{\mathrm{KL}}\left[\Pr\big(X_{1\ldots n}|Y\big)\middle\|\prod_{i=1}^{n}\Pr\big(X_i|Y\big)\right] - \mathrm{D}_{\mathrm{KL}}\left[\Pr(X_{1\ldots n})\middle\|\Pr_{\mathrm{ind}}(\mathbf{X})\right]\\
&= \mathrm{TC}\left(X_1;\cdots;X_n|Y\right) - \mathrm{D}_{\mathrm{KL}}\left[\Pr(X_{1\ldots n})\middle\|\Pr_{\mathrm{ind}}(\mathbf{X})\right]\ . && (2.23)
\end{aligned}
$$

where $\mathrm{TC}\left(X_1;\cdots;X_n|Y\right)$ is the conditional total correlation among the predictors given Y.

Part II

Making Progress

Chapter 3

First Nontrivial Lowerbound on Synergy

Remark: This chapter borrows liberally from the joint paper [14].

3.1 Introduction

Introduced in [36], *Partial Information Decomposition* (PID) is an immensely useful framework for deepening our understanding of multivariate interactions, particularly our understanding of informational redundancy and synergy. To harness the PID framework, the user brings her own measure of *intersection information*, $\mathrm{I}_\cap(X_1, \ldots, X_n : Y)$, which quantifies the magnitude of information that each of the n predictors $X_1, \ldots, X_n$ conveys "redundantly" about a target random variable Y. An antichain lattice of redundant, unique, and synergistic partial informations is built from the intersection information.

In [36], the authors propose to use the following quantity, $\mathrm{I}_{\min}$, as the intersection information measure:

$$\begin{aligned} \mathrm{I}_{\min}\left(X_1, \ldots, X_n : Y\right) &\equiv \sum_y \Pr(y) \min_i \mathrm{I}(X_i : Y = y) \\ &= \sum_y \Pr(y) \min_i \mathrm{D}_{\mathrm{KL}}\left[\Pr\big(X_i|y\big) \middle\| \Pr(X_i)\right] , \end{aligned} \tag{3.1}$$

where D_{KL} is the Kullback-Leibler divergence.

Though $\mathrm{I}_{\min}$ is an intuitive and plausible choice for the intersection information, [15] showed that $\mathrm{I}_{\min}$ has counterintuitive properties. In particular, $\mathrm{I}_{\min}$ calculates one bit of redundant information for example UNQ (Figure 3.3). It does this because each input shares one bit of information with the output. However, it is quite clear that the shared informations are, in fact, different: X_1 provides the low bit, while X_2 provides the high bit. This led to the conclusion that $\mathrm{I}_{\min}$ *over-estimates* the ideal intersection information measure by focusing only on *how much* information the inputs

provide to the output. An ideal measure of intersection information must recognize that there are non-equivalent ways of providing information to the output. The search for an improved intersection information measure ensued, continued through [18, 7, 23], and today a widely accepted intersection information measure remains undiscovered.

Here we do not definitively solve this problem, but we present a strong candidate intersection information measure for the special case of *zero-error* information. This is useful in of itself because it provides a template for how the yet undiscovered ideal intersection information measure for Shannon mutual information could work. Alternatively, if a Shannon intersection information measure with the same properties does not exist, then we have learned something significant.

In the next section, we introduce some definitions, some notation, and a necessary lemma. We also extend and clarify the desired properties for intersection information. In Section 3.4 we introduce zero-error information and its intersection information measure. In Section 3.5 we use the same methodology to produce a novel candidate for the Shannon intersection information. In Section 3.6 we show the successes and shortcomings of our candidate intersection information measure using example circuits. Finally, in Section 3.8 we summarize our progress towards the ideal intersection information measure and suggest directions for improvement. The Appendix is devoted to technical lemmas and their proofs, to which we refer in the main text.

3.2 Two examples elucidating desired properties for synergy

To help the reader develop intuition for any proper measure of synergy, we illustrate some desired properties of synergistic information with pedagogical examples. Both examples are derived from example Xor.

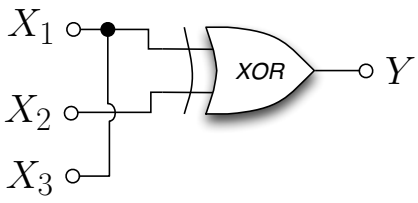
3.2.1 XorDuplicate: Synergy is invariant to duplicating a predictor

Example XorDuplicate (Figure 3.1) adds a third predictor, X_3, a copy of predictor X_1, to Xor. Whereas in Xor the target Y is specified only by coalition X_1X_2, duplicating predictor X_1 as X_3 makes the target equally specifiable by coalition X_3X_2.

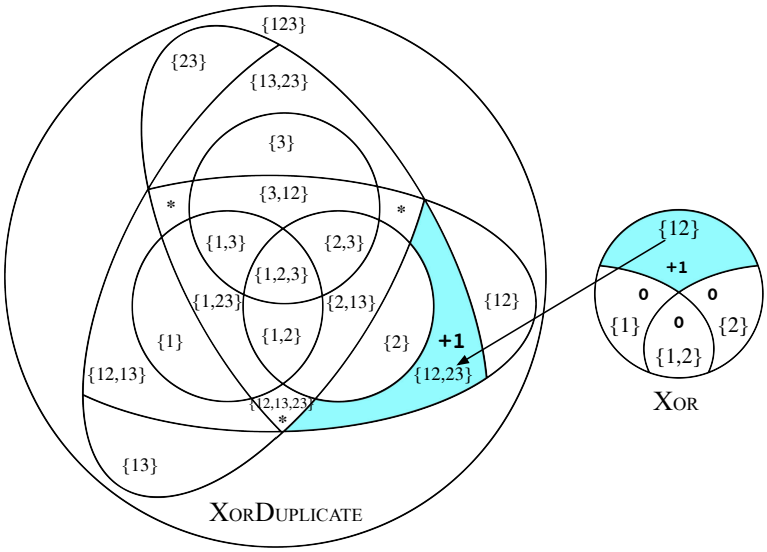
Although now two different coalitions identically specify Y, mutual information is invariant to duplicates, e.g. $\mathrm{I}(X_1X_2X_3 : Y) = \mathrm{I}(X_1X_2 : Y)$ bit. For synergistic information to be likewise bounded between zero and the total mutual information $\mathrm{I}(X_{1\ldots n} : Y)$, synergistic information must similarly be invariant to duplicates, e.g. the synergistic information between set $\{X_1, X_2\}$ and Y must be the same as the synergistic information between $\{X_1, X_2, X_3\}$ and Y. This makes sense because if synergistic information is defined as the information in the whole beyond its parts, duplicating a part does not increase the net information provided by the parts. Altogether, we assert that *duplicating a predictor does not change the synergistic information.*

X_1	X_2	X_3	Y	
0	0	0	0	1/4
0	1	0	1	1/4
1	0	1	1	1/4
1	1	1	0	1/4

(a) $\Pr(x_1, x_2, x_3, y)$



(b) circuit diagram



(c) PI-diagram

Figure 3.1: Example XorDuplicate shows that duplicating predictor X_1 as X_3 turns the single-coalition synergy $\{12\}$ into the multi-coalition synergy $\{12, 23\}$. After duplicating X_1, the coalition X_3X_2 as well as coalition X_1X_2 specifies Y. Synergistic information is unchanged from Xor, $\mathrm{I}(X_3X_2 : Y) = \mathrm{I}(X_1X_2 : Y) = \mathrm{H}(Y) = 1$ bit.

3.2.2 XorLoses: Adding a new predictor can decrease synergy

Example XorLoses (Figure 3.2) adds a third predictor, X_3, to Xor and concretizes the distinction between synergy and "redundant synergy". In XorLoses the target Y has one bit of uncertainty, and just as in example Xor the coalition X_1X_2 fully specifies the target, $\mathrm{I}(X_1X_2 : Y) = \mathrm{H}(Y) = 1$ bit. However, XorLoses has *zero* intuitive synergy because the newly added singleton predictor, X_3, fully specifies Y by itself. This makes the synergy between X_1 and X_2 *completely redundant*—everything the coalition X_1X_2 specifies is now already specified by the singleton X_3.

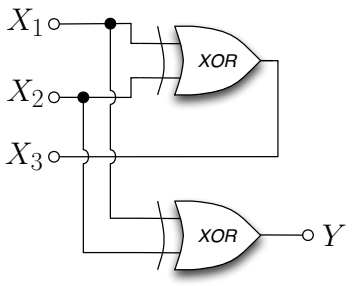
3.3 Preliminaries

3.3.1 Informational Partial Order and Equivalence

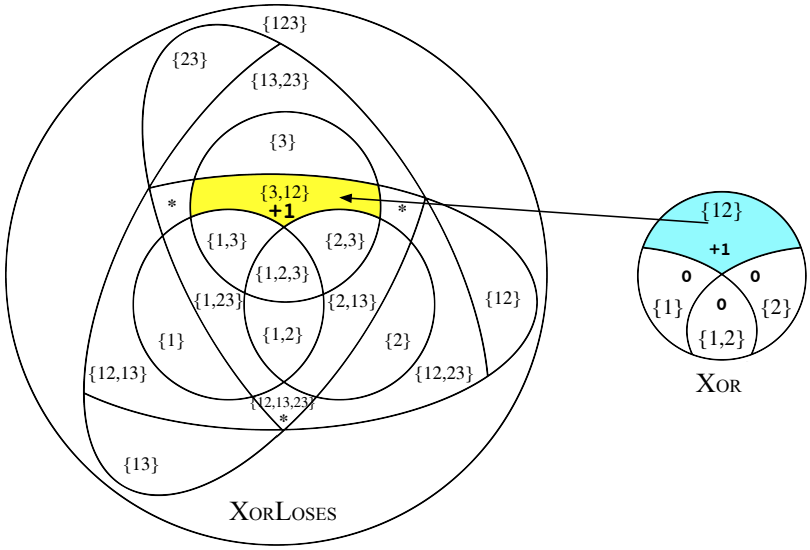
We assume an underlying probability space on which we define random variables, as denoted by capital letters (e.g., X, Y, and Z). In this paper, we consider only random variables taking values

X_1	X_2	X_3	Y	
0	0	0	0	1/4
0	1	1	1	1/4
1	0	1	1	1/4
1	1	0	0	1/4

(a) $\Pr(x_1, x_2, x_3, y)$



(b) circuit diagram



(c) PI-diagram

Figure 3.2: Example XorLoses. Target Y is fully specified by the coalition X_1X_2 as well as by the singleton X_3. $\mathrm{I}(X_1X_2 : Y) = \mathrm{I}(X_3 : Y) = \mathrm{H}(Y) = 1$ bit. Therefore, the information synergistically specified by coalition X_1X_2 is a redundant synergy.

on finite spaces.

Given random variables X and Y, we write $X \preceq Y$ to signify that there exists a measurable function f such that $X = f(Y)$. In this case, following the terminology in [22], we say that X is *informationally poorer* than Y; this induces a partial order on the set of random variables. Similarly, we write $X \succeq Y$ if $Y \preceq X$, in which case we say X is *informationally richer* than Y.

If X and Y are such that $X \preceq Y$ and $X \succeq Y$, then we write $X \cong Y$. In this case, again following [22], we say that X and Y are *informationally equivalent*. In other words, $X \cong Y$ if and only if there's an invertible function between X and Y, i.e., one can relabel the values of X to obtain a random value that is equal to Y and vice versa.

This "information-equivalence" relation can easily be shown to be an equivalence relation, so that we can partition the set of all random variables into disjoint equivalence classes. The $\preceq$ ordering is invariant within these equivalence classes in the following sense: if $X \preceq Y$ and $Y \cong Z$, then $X \preceq Z$. Similarly, if $X \preceq Y$ and $X \cong Z$, then $Z \preceq Y$. Moreover, within each equivalence class, the entropy

is invariant, as stated formally in Lemma 3.3.1 below.

3.3.2 Information Lattice

Next, we follow [22] and consider the *join* and *meet* operators. These operators were defined for *information elements*, which are σ-algebras, or, equivalently, equivalence classes of random variables. We deviate from [22], though, by defining the join and meet operators for random variables, but we do preserve their conceptual properties.

Given random variables X and Y, we define $X \curlyvee Y$ (called the *join* of X and Y) to be an informationally poorest ("smallest" in the sense of the partial order $\preceq$) random variable such that $X \preceq X \curlyvee Y$ and $Y \preceq X \curlyvee Y$. In other words, if Z is such that $X \preceq Z$ and $Y \preceq Z$, then $X \curlyvee Y \preceq Z$. Note that $X \curlyvee Y$ is unique only up to equivalence with respect to $\cong$. In other words, $X \curlyvee Y$ does not define a specific, unique random variable. Nonetheless, standard information-theoretic quantities are invariant over the set of random variables satisfying the condition specified above. For example, the entropy of $X \curlyvee Y$ is invariant over the entire equivalence class of random variables satisfying the condition above (by Lemma 3.3.1(a) below). Similarly, the inequality $Z \preceq X \curlyvee Y$ does not depend on the specific random variable chosen, as long as it satisfies the condition above. Note that the pair (X, Y) is an instance of $X \curlyvee Y$.

In a similar vein, given random variables X and Y, we define $X \curlywedge Y$ (called the *meet* of X and Y) to be an informationally richest random variable ("largest" in the sense of $\succeq$) such that $X \curlywedge Y \preceq X$ and $X \curlywedge Y \preceq Y$. In other words, if Z is such that $Z \preceq X$ and $Z \preceq Y$, then $Z \preceq X \curlywedge Y$. Following [16], we also call $X \curlywedge Y$ the *common random variable* of X and Y. Again, considering the entropy of $X \curlywedge Y$ or the inequality $Z \preceq X \curlywedge Y$ does not depend on the specific random variable chosen, as long as it satisfies the condition above.

The $\curlyvee$ and $\curlywedge$ operators satisfy the algebraic properties of a *lattice* [22]. In particular, the following hold:

- commutative laws: $X \curlyvee Y \cong Y \curlyvee X$ and $X \curlywedge Y \cong Y \curlywedge X$
- associative laws: $X \curlyvee (Y \curlyvee Z) \cong (X \curlyvee Y) \curlyvee Z$ and $X \curlywedge (Y \curlywedge Z) \cong (X \curlywedge Y) \curlywedge Z$
- absorption laws: $X \curlyvee (X \curlywedge Y) \cong X$ and $X \curlywedge (X \curlyvee Y) \cong X$)
- idempotent laws: $X \curlyvee X \cong X$ and $X \curlywedge X \cong X$
- generalized absorption laws: if $X \preceq Y$, then $X \curlyvee Y \cong Y$and $X \curlywedge Y \cong X$.

Finally, the partial order $\preceq$ is preserved under $\curlyvee$ and $\curlywedge$, i.e., if $X \preceq Y$, then $X \curlyvee Z \preceq Y \curlyvee Z$ and $X \curlywedge Z \preceq X \curlywedge Z$.

3.3.3 Invariance and Monotonicity of Entropy

Let $\mathrm{H}(\cdot)$ represent the entropy function, and $\mathrm{H}\big(\cdot|\cdot\big)$ the conditional entropy. To be consistent with the colon in the intersection information, we denote the Shannon mutual information between X and Y by $\mathrm{I}(X:Y)$ instead of the more common $\mathrm{I}(X;Y)$. Lemma 3.3.1 establishes the invariance and monotonicity of the entropy and conditional entropy functions with respect to $\cong$ and $\preceq$.

Lemma 3.3.1. *The following hold:*

(a) If $X \cong Y$*, then* $\mathrm{H}(X) = \mathrm{H}(Y)$*,* $\mathrm{H}(X|Z) = \mathrm{H}(Y|Z)$*, and* $\mathrm{H}(Z|X) = \mathrm{H}(Z|Y)$*.*

(b) If $X \preceq Y$*, then* $\mathrm{H}(X) \leq \mathrm{H}(Y)$*,* $\mathrm{H}(X|Z) \leq \mathrm{H}(Y|Z)$*, and* $\mathrm{H}(Z|X) \geq \mathrm{H}(Z|Y)$*.*

(c) $X \preceq Y$ *if and only if* $\mathrm{H}(X|Y) = 0$*.*

Proof. Part (a) follows from [22], Proposition 1. Part (c) follows from [22], Proposition 4. The first two desired inequalities in part (b) follow from [22], Proposition 5. Now we show that if $X \preceq Y$, then $\mathrm{H}\big(Z|X\big) \geq \mathrm{H}\big(Z|Y\big)$. Suppose that $X \preceq Y$. Then, by the generalized absorption law, $X \curlyvee Y \cong Y$. We have

$$\begin{aligned}
\mathrm{I}(Z:Y) &= \mathrm{H}(Y) - \mathrm{H}(Y|Z) \\
&= \mathrm{H}(X \curlyvee Y) - \mathrm{H}(X \curlyvee Y|Z) \quad \text{by part (a)} \\
&= \mathrm{I}(Z:X \curlyvee Y) \\
&= \mathrm{I}(Z:X) + \mathrm{I}(Z:Y|X) \\
&\geq \mathrm{I}(Z:X)\,.
\end{aligned}$$

Substituting $\mathrm{I}(Z:Y) = \mathrm{H}(Z) - \mathrm{H}(Z|Y)$ and $\mathrm{I}(Z:X) = \mathrm{H}(Z) - \mathrm{H}(Z|X)$, we obtain $\mathrm{H}(Z|X) \geq \mathrm{H}(Z|Y)$ as desired. □

Remark: Because $(X,Y) \cong X \curlyvee Y$ as noted before, we also have $\mathrm{H}(X,Y) = \mathrm{H}(X \curlyvee Y)$ by Lemma 3.3.1(a).

3.3.4 Desired Properties of Intersection Information

There are currently 12 intuitive properties that we wish the ideal intersection information measure $\mathrm{I}_\cap$ to satisfy. Some are new (e.g. $(\mathbf{M_1})$, (**Eq**), (**LB**)), but most were introduced earlier, in various forms, Refs. [36, 15, 18, 7, 23]. They are as follows:

(**GP**) Global Positivity: $\mathrm{I}_\cap(X_1, \ldots, X_n : Y) \geq 0$, and $\mathrm{I}_\cap(X_1, \ldots, X_n : Y) = 0$ if Y is a constant.

(**Eq**) Equivalence-Class Invariance: $\mathrm{I}_\cap(X_1, \ldots, X_n : Y)$ is invariant under substitution of X_i (for any $i = 1, \ldots, n$) or Y by an informationally equivalent random variable.

(**TM**) Target Monotonicity: If $Y \preceq Z$, then $\mathrm{I}_\cap(X_1, \ldots, X_n : Y) \leq \mathrm{I}_\cap(X_1, \ldots, X_n : Z)$.

(**$\mathbf{M_0}$**) Weak Monotonicity: $\mathrm{I}_\cap(X_1, \ldots, X_n, W : Y) \leq \mathrm{I}_\cap(X_1, \ldots, X_n : Y)$ with equality if there exists $Z \in \{X_1, \ldots, X_n\}$ such that $Z \preceq W$.

(**$\mathbf{S_0}$**) Weak Symmetry: $\mathrm{I}_\cap(X_1, \ldots, X_n : Y)$ is invariant under reordering of $X_1, \ldots, X_n$.

Remark: If (**$\mathbf{S_0}$**) is satisfied, the first argument of $\mathrm{I}_\cap(X_1, \ldots, X_n : Y)$ can be treated as a *set* of random variables rather than a *list*. In this case, the notation $\mathrm{I}_\cap\big(\{X_1, \ldots, X_n\} : Y\big)$ would also be appropriate.

For the next set of properties, $\mathcal{I}(X : Y)$ is a given normative measure of information between X and Y. For example, $\mathcal{I}(X : Y)$ could denote the Shannon mutual information; i.e., $\mathcal{I}(X : Y) = \mathrm{I}(X : Y)$. Alternatively, as discussed in the next section, we might take $\mathcal{I}(X : Y)$ to be the zero-error information. Yet other possibilities for $\mathcal{I}(X : Y)$ include the Wyner common information [38] or the quantum mutual information [8]. The following are desired properties of intersection information *relative to* the given information measure $\mathcal{I}$.

(**LB**) Lowerbound: If $Q \preceq X_i$ for all $i = 1, \ldots, n$, then $\mathrm{I}_\cap(X_1, \ldots, X_n : Y) \geq \mathcal{I}(Q : Y)$. Under a mild assumption,[1] this equates to $\mathrm{I}_\cap(X_1, \ldots, X_n : Y) \;\geq\; \mathcal{I}(X_1 \curlywedge \cdots \curlywedge X_n : Y)$.

(**SR**) Self-Redundancy: $\mathrm{I}_\cap(X_1 : Y) = \mathcal{I}(X_1 : Y)$. The intersection information a single predictor X_1 conveys about the target Y is equal to the information between the predictor and the target given by the information measure $\mathcal{I}$.

(**Id**) Identity: $\mathrm{I}_\cap(X, Y : X \curlyvee Y) = \mathcal{I}(X : Y)$.

(**$\mathbf{LP_0}$**) Weak Local Positivity: $\mathrm{I}_\cap(X_1, X_2 : Y) \geq \mathcal{I}(X_1 : Y) + \mathcal{I}(X_2 : Y) - \mathcal{I}(X_1 \curlyvee X_2 : Y)$. In other words, for $n = 2$ predictors, the derived "partial informations" defined in [36] are nonnegative when both (**$\mathbf{LP_0}$**) and (**GP**) hold.

Finally, we have the less obvious "strong" properties.

(**$\mathbf{M_1}$**) Strong Monotonicity: $\mathrm{I}_\cap(X_1, \ldots, X_n, W : Y) \leq \mathrm{I}_\cap(X_1, \ldots, X_n : Y)$ with equality if there exists $Z \in \{X_1, \ldots, X_n, Y\}$ such that $Z \preceq W$.

(**$\mathbf{S_1}$**) Strong Symmetry: $\mathrm{I}_\cap(X_1, \ldots, X_n : Y)$ is invariant under reordering of $X_1, \ldots, X_n, Y$.

(**$\mathbf{LP_1}$**) Strong Local Positivity: For all n, the derived "partial informations" defined in [36] are nonnegative.

[1] See Lemmas 3.C.1 and 3.C.2 in Appendix 3.C.1.

Properties (**Eq**), (**LB**), and ($\mathbf{M_1}$) are novel and are introduced for the first time here. Given $\mathrm{I}_\cap$, $X_1, \ldots, X_n$, Y, and Z, we define the conditional $\mathrm{I}_\cap$ as:

$$\mathrm{I}_\cap(X_1, \ldots, X_n : Z|Y) \equiv \mathrm{I}_\cap(X_1, \ldots, X_n : Y \curlyvee Z) - \mathrm{I}_\cap(X_1, \ldots, X_n : Y) \ .$$

This definition of $\mathrm{I}_\cap(X_1, \ldots, X_n : Z|Y)$ gives rise to the familiar "chain rule":

$$\mathrm{I}_\cap(X_1, \ldots, X_n : Y \curlyvee Z) = \mathrm{I}_\cap(X_1, \ldots, X_n : Y) + \mathrm{I}_\cap(X_1, \ldots, X_n : Z|Y) \ .$$

Some provable[2] properties are:

- $\mathrm{I}_\cap(X_1, \ldots, X_n : Z|Y) \geq 0$.
- $\mathrm{I}_\cap(X_1, \ldots, X_n : Z|Y) = \mathrm{I}_\cap(X_1, \ldots, X_n : Z)$ if Y is a constant.

3.4 Candidate Intersection Information for Zero-Error Information

3.4.1 Zero-Error Information

Introduced in [37], the *zero-error information*, or *Gács-Körner common information*, is a stricter variant of Shannon mutual information. Whereas the mutual information $\mathrm{I}(A:B)$ quantifies the magnitude of information A conveys about B with an arbitrarily small error $\epsilon > 0$, the zero-error information, denoted $\mathrm{I}^0(A:B)$, quantifies the magnitude of information A conveys about B with *exactly zero* error, i.e., $\epsilon = 0$. The zero-error information between A and B equals the entropy of the *common random variable* $A \curlywedge B$,

$$\mathrm{I}^0(A:B) \equiv \mathrm{H}(A \curlywedge B) \ .$$

An algorithm for computing an instance of the common random variable between two random variables is provided in [37], and straightforwardly generalizes to n random variables.[3]

Zero-error information has several notable properties, but the most salient is that it is nonnegative and bounded by the mutual information,

$$0 \leq \mathrm{I}^0(A:B) \leq \mathrm{I}(A:B) \, .$$

[2] See Lemma 3.C.3 in Appendix 3.C.1.
[3] See Appendix 3.A.

This generalizes to arbitrary n:

$$0 \leq \mathrm{I}^0(X_1 : \cdots : X_n) \leq \min_{i,j} \mathrm{I}\big(X_i : X_j\big) \,.$$

3.4.2 Intersection Information for Zero-Error Information

It is pleasingly straightforward to define a palatable intersection information for zero-error information (i.e., setting $\mathcal{I} = \mathrm{I}^0$ as the normative measure of information). We propose the zero-error intersection information, $\mathrm{I}^0_\curlywedge(X_1, \ldots, X_n : Y)$, as the maximum zero-error information $\mathrm{I}^0(Q : Y)$ that some random variable Q conveys about Y, subject to Q being a function of each predictor $X_1, \ldots, X_n$:

$$\begin{aligned} \mathrm{I}^0_\curlywedge(X_1, \ldots, X_n : Y) \equiv \max_{\Pr(Q|Y)} & \; \mathrm{I}^0(Q : Y) \\ & \text{subject to } \forall i \in \{1, \ldots, n\} : Q \preceq X_i \,. \end{aligned} \tag{3.2}$$

Basic algebra[4] shows that a maximizing Q is the common random variable across all predictors. This substantially simplifies eq. (3.2) to:

$$\begin{aligned} \mathrm{I}^0_\curlywedge(X_1, \ldots, X_n : Y) &= \mathrm{I}^0(X_1 \curlywedge \cdots \curlywedge X_n : Y) \\ &= \mathrm{H}\big[(X_1 \curlywedge \cdots \curlywedge X_n) \curlywedge Y\big] \\ &= \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y) \,. \end{aligned} \tag{3.3}$$

Importantly, the zero-error information, $\mathrm{I}^0_\curlywedge(X_1, \ldots, X_n : Y)$ satisfies ten of the twelve desired properties from Section 3.3.4, leaving only ($\mathbf{LP_0}$) and ($\mathbf{LP_1}$) unsatisfied.[5]

3.5 Candidate Intersection Information for Shannon Information

In the last section, we defined an intersection information for zero-error information which satisfies the vast majority of desired properties. This is a solid start, but an intersection information for Shannon mutual information remains the goal. Towards this end, we use the same method as in eq. (3.2), leading to $\mathrm{I}_\curlywedge$, our candidate intersection information measure for Shannon mutual information,

$$\begin{aligned} \mathrm{I}_\curlywedge(X_1, \ldots, X_n : Y) \equiv \max_{\Pr(Q|Y)} & \; \mathrm{I}(Q : Y) \\ & \text{subject to } Q \preceq X_i \;\forall i \in \{1, \ldots, n\} \,. \end{aligned} \tag{3.4}$$

[4] See Lemma 3.D.1 in Appendix 5.D.
[5] See Lemmas 3.C.4, 3.C.5, 3.C.6 in Appendix 3.C.2.

With some algebra[6] this similarly simplifies to,

$$\mathrm{I}_{\curlywedge}(X_1, \ldots, X_n : Y) = \mathrm{I}(X_1 \curlywedge \cdots \curlywedge X_n : Y) \ . \tag{3.5}$$

Unfortunately $\mathrm{I}_{\curlywedge}$ does not satisfy as many of the desired properties as $\mathrm{I}^0_{\curlywedge}$. However, our candidate $\mathrm{I}_{\curlywedge}$ still satisfies 7 of the 12 properties (Table 3.1), most importantly the enviable (**TM**),[7] which has, until now, not been satisfied by any proposed measure. Table 3.1 lists the desired properties satisfied by $\mathrm{I}_{\min}$, $\mathrm{I}_{\curlywedge}$, and $\mathrm{I}^0_{\curlywedge}$. For reference, we also include $\mathrm{I}_{\mathrm{red}}$, the proposed measure from [18].

Comparing the three subject intersection information measures,[8] we have:

$$0 \leq \mathrm{I}^0_{\curlywedge}(X_1, \ldots, X_n : Y) \leq \mathrm{I}_{\curlywedge}(X_1, \ldots, X_n : Y) \leq \mathrm{I}_{\min}(X_1, \ldots, X_n : Y) \ . \tag{3.6}$$

Property		$\mathrm{I}_{\min}$	$\mathrm{I}_{\mathrm{red}}$	$\mathrm{I}_{\curlywedge}$	$\mathrm{I}^0_{\curlywedge}$
(**GP**)	Global Positivity	✓	✓	✓	✓
(**Eq**)	Equivalence-Class Invariance	✓	✓	✓	✓
(**TM**)	Target Monotonicity			✓	✓
($\mathbf{M_0}$)	Weak Monotonicity	✓		✓	✓
($\mathbf{S_0}$)	Weak Symmetry	✓	✓	✓	✓
(**LB**)	Lowerbound	✓	✓	✓	✓
(**SR**)	Self-Redundancy	✓	✓	✓	✓
(**Id**)	Identity		✓		✓
($\mathbf{LP_0}$)	Weak Local Positivity	✓	✓		
($\mathbf{M_1}$)	Strong Monotonicity				✓
($\mathbf{S_1}$)	Strong Symmetry				✓
($\mathbf{LP_1}$)	Strong Local Positivity	✓			

Table 3.1: The $\mathrm{I}_\cap$ desired properties each measure satisfies.

Despite not satisfying ($\mathbf{LP_0}$), $\mathrm{I}_{\curlywedge}$ remains an important stepping-stone towards the ideal Shannon $\mathrm{I}_\cap$. First, $\mathrm{I}_{\curlywedge}$ captures what is inarguably redundant information (the common random variable); this makes $\mathrm{I}_{\curlywedge}$ necessarily a lower bound on any reasonable redundancy measure. Second, it is the first proposal to satisfy target monotonicity and the associated chain rule. Lastly, $\mathrm{I}_{\curlywedge}$ is the first measure to reach intuitive answers in many canonical situations, while also being generalizable to an arbitrary number of inputs.

[6] See Lemma 3.D.2 in Appendix 5.D.
[7] See Lemmas 3.C.7, 3.C.8, 3.C.9 in Appendix 3.C.3.
[8] See Lemma 3.D.3 in Appendix 5.D.

3.6 Three Examples Comparing $\mathrm{I}_{\min}$ and $\mathrm{I}_\curlywedge$

Examples Unq and RdnXor illustrate $\mathrm{I}_\curlywedge$'s successes, and example ImperfectRdn illustrates $\mathrm{I}_\curlywedge$'s paramount deficiency. For each example we show the joint distribution $\Pr(x_1, x_2, y)$, a diagram, and the decomposition derived from setting $\mathrm{I}_{\min}$ / $\mathrm{I}_\curlywedge$ as the $\mathrm{I}_\cap$ measure. At each lattice junction, the left number is the $\mathrm{I}_\cap$ value of that node, and the number in parentheses is the I_∂ value.[9] Readers unfamiliar with the $n = 2$ partial information lattice should consult [36], but in short, I_∂ measures the amount of "new" information at this node in the lattice compared to nodes lower in the lattice. Except for ImperfectRdn, measures $\mathrm{I}_\curlywedge$ and $\mathrm{I}^0_\curlywedge$ reach the same decomposition for all presented examples. Per [36], the four partial informations are calculated as follows:

$$\begin{aligned}
\mathrm{I}_\partial(X_1, X_2 : Y) &= \mathrm{I}_\cap(X_1, X_2 : Y) \\
\mathrm{I}_\partial(X_1 : Y) &= \mathrm{I}(X_1 : Y) - \mathrm{I}_\cap(X_1, X_2 : Y) \\
\mathrm{I}_\partial(X_2 : Y) &= \mathrm{I}(X_2 : Y) - \mathrm{I}_\cap(X_1, X_2 : Y) \\
\mathrm{I}_\partial(X_1 \curlyvee X_2 : Y) &= \mathrm{I}(X_1 \curlyvee X_2 : Y) - \mathrm{I}(X_1 : Y) - \mathrm{I}(X_2 : Y) + \mathrm{I}_\cap(X_1, X_2 : Y) \\
&= \mathrm{I}(X_1 \curlyvee X_2 : Y) - \mathrm{I}_\partial\left(X_1 : Y\right) - \mathrm{I}_\partial\left(X_2 : Y\right) - \mathrm{I}_\partial\left(X_1, X_2 : Y\right) \ .
\end{aligned} \tag{3.7}$$

Example Unq (Figure 3.3). The desired decomposition for this example is two bits of unique information; X_1 uniquely specifies one bit of Y, and X_2 uniquely specifies the other bit of Y. The chief criticism of $\mathrm{I}_{\min}$ in [15] was that $\mathrm{I}_{\min}$ calculated one bit of redundancy and one bit of synergy for Unq (Figure 3.3c). We see that unlike $\mathrm{I}_{\min}$, $\mathrm{I}_\curlywedge$ satisfyingly arrives at two bits of unique information. This is easily seen by the inequality,

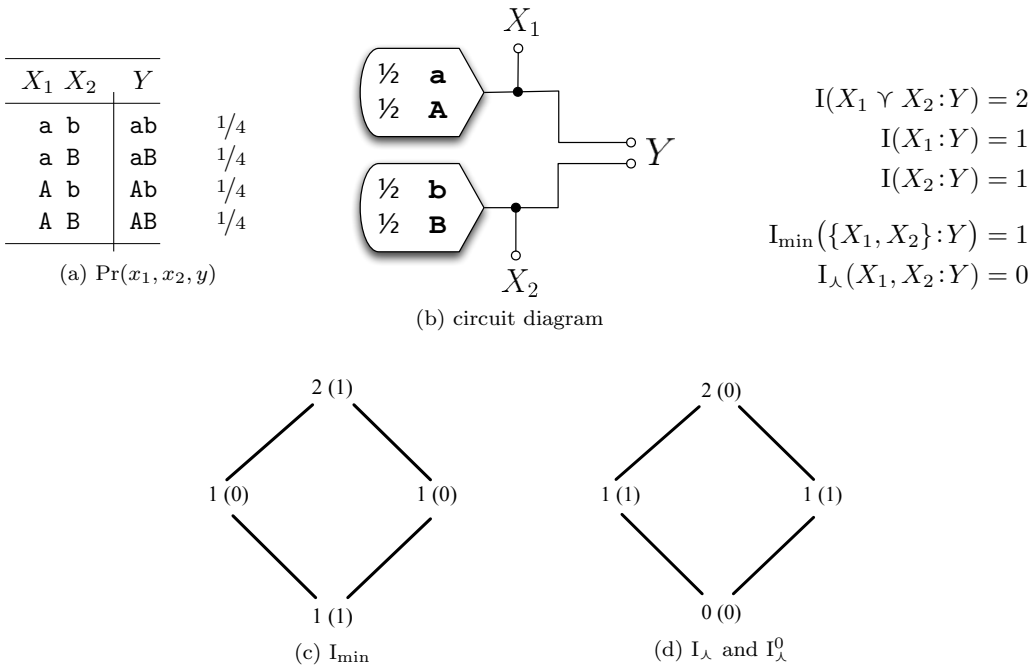
$$0 \leq \mathrm{I}_\curlywedge(X_1, X_2 : Y) \leq \mathrm{H}(X_1 \curlywedge X_2) \leq \mathrm{I}(X_1 : X_2) = 0 \text{ bits} \ . \tag{3.8}$$

Therefore, as $\mathrm{I}(X_1 : X_2) = 0$, we have $\mathrm{I}_\curlywedge(X_1, X_2 : Y) = 0$ bits leading to $\mathrm{I}_\partial(X_1 : Y) = 1$ bit and $\mathrm{I}_\partial(X_2 : Y) = 1$ bit (Figure 3.3d).

Example RdnXor (Figure 3.4). In [15], RdnXor was an example where $\mathrm{I}_{\min}$ shined by reaching the desired decomposition of one bit of redundancy and one bit of synergy. We see that $\mathrm{I}_\curlywedge$ finds this same answer. $\mathrm{I}_\curlywedge$ extracts the common random variable within X_1 and X_2, the `r`/`R` bit, and calculates the mutual information between the common random variable and Y to arrive at $\mathrm{I}_\curlywedge(X_1, X_2 : Y) = 1$ bit.

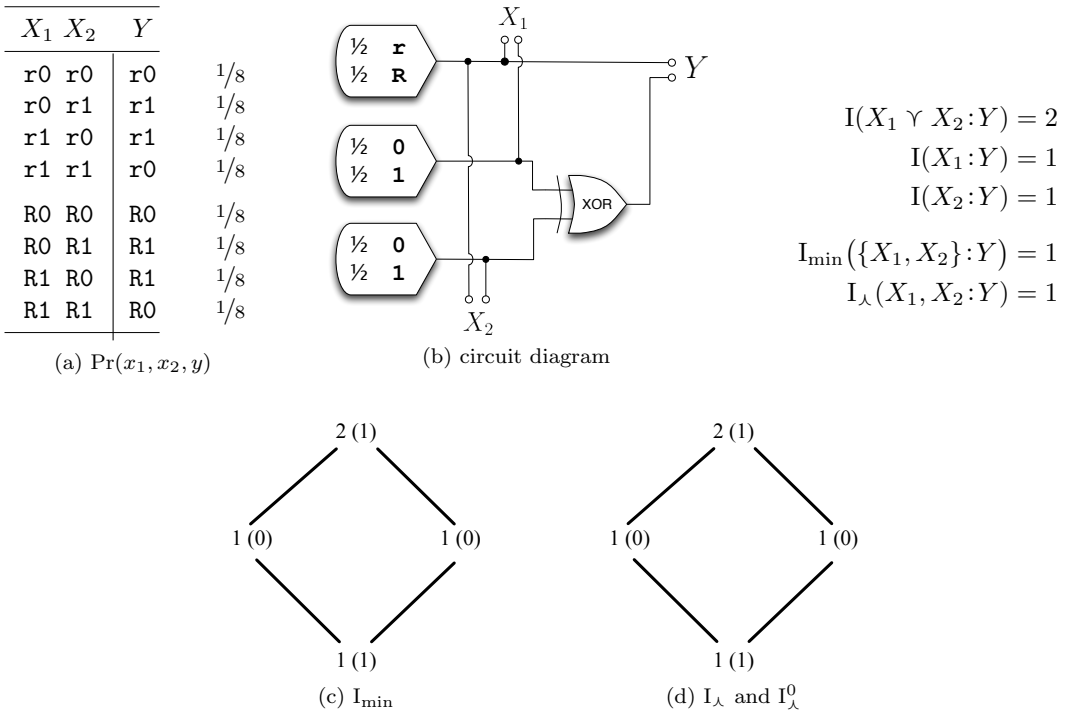
Example ImperfectRdn (Figure 3.5). ImperfectRdn highlights the foremost shortcoming of $\mathrm{I}_\curlywedge$; $\mathrm{I}_\curlywedge$ does not detect "imperfect" or "lossy" correlations between X_1 and X_2. Given ($\mathbf{LP_0}$),

[9] This is the same notation used in [7].



(a) $\Pr(x_1, x_2, y)$ (b) circuit diagram (c) $\mathrm{I}_{\min}$ (d) $\mathrm{I}_\wedge$ and $\mathrm{I}^0_\wedge$

Figure 3.3: Example UNQ. This is the canonical example of unique information. X_1 and X_2 each uniquely specify a single bit of Y. This is the simplest example where $\mathrm{I}_{\min}$ calculates an undesirable decomposition (c) of one bit of redundancy and one bit of synergy. $\mathrm{I}_\wedge$ and $\mathrm{I}^0_\wedge$ each calculate the desired decomposition (d).



(a) $\Pr(x_1, x_2, y)$ (b) circuit diagram (c) $\mathrm{I}_{\min}$ (d) $\mathrm{I}_\wedge$ and $\mathrm{I}^0_\wedge$

Figure 3.4: Example RDNXOR. This is the canonical example of redundancy and synergy coexisting. $\mathrm{I}_{\min}$ and $\mathrm{I}_\wedge$ each reach the desired decomposition of one bit of redundancy and one bit of synergy. This is the simplest example demonstrating $\mathrm{I}_\wedge$ and $\mathrm{I}^0_\wedge$ correctly extracting the embedded redundant bit within X_1 and X_2.

we can determine the desired decomposition analytically. First, $\mathrm{I}(X_1 \curlyvee X_2 : Y) = \mathrm{I}(X_1 : Y) = 1$ bit; therefore, $\mathrm{I}\big(X_2 : Y | X_1\big) = \mathrm{I}(X_1 \curlyvee X_2 : Y) - \mathrm{I}(X_1 : Y) \;\; = \;\; 0$ bits. This determines two of the partial informations—the synergistic information $\mathrm{I}_\partial(X_1 \curlyvee X_2 : Y)$ and the unique information $\mathrm{I}_\partial(X_2 : Y)$ are both zero. Then, the redundant information $\mathrm{I}_\partial(X_1, X_2 : Y) = \mathrm{I}(X_2 : Y) - \mathrm{I}_\partial(X_2 : Y) = \mathrm{I}(X_2 : Y) = 0.99$ bits. Having determined three of the partial informations, we compute the final unique information $\mathrm{I}_\partial(X_1 : Y) = \mathrm{I}(X_1 : Y) - 0.99 = 0.01$ bits.

How well do $\mathrm{I}_{\min}$ and $\mathrm{I}_\curlywedge$ match the desired decomposition of ImperfectRdn? We see that $\mathrm{I}_{\min}$ calculates the desired decomposition (Figure 3.5c); however, $\mathrm{I}_\curlywedge$ does not (Figure 3.5d). Instead, $\mathrm{I}_\curlywedge$ calculates zero redundant information, that $\mathrm{I}_\cap(X_1, X_2 : Y) = 0$ bits. This unpleasant answer arises from $\Pr(X_1 = \mathtt{1}, X_2 = \mathtt{0}) > 0$. If this were zero, ImperfectRdn reverts to the example Rdn (Figure **??** in Appendix 3.E) where both $\mathrm{I}_\curlywedge$ and $\mathrm{I}_{\min}$ reach the desired one bit of redundant information. Due to the nature of the common random variable, $\mathrm{I}_\curlywedge$ only sees the "deterministic" correlations between X_1 and X_2—add even an iota of noise between X_1 and X_2 and $\mathrm{I}_\curlywedge$ plummets to zero. This highlights a related issue with $\mathrm{I}_\curlywedge$—it is not continuous; an arbitrarily small change in the probability distribution can result in a discontinuous jump in the value of $\mathrm{I}_\curlywedge$. As with traditional information measures, such as the entropy and the mutual information, it may be desirable to have an $\mathrm{I}_\cap$ measure that is continuous over the simplex.

To summarize, ImperfectRdn shows that when there are additional "imperfect" correlations between A and B, i.e. $\mathrm{I}(A : B | A \curlywedge B) > 0$, $\mathrm{I}_\curlywedge$ sometimes *underestimates* the ideal $\mathrm{I}_\cap(A, B : Y)$.

3.7 Negative synergy and state-dependent (GP)

In ImperfectRdn we saw $\mathrm{I}_\curlywedge$ calculate a synergy of -0.99 bits (Figure 3.5d). What does this mean? Could negative synergy be a "real" property of Shannon information? When $n = 2$, it's fairly easy to diagnose the cause of negative synergy from the equation for $\mathrm{I}_\partial(X_1, X_2 : Y)$ in eq. (3.7). Given (**GP**) and (**SR**), negative synergy occurs if and only if,

$$\begin{aligned} \mathrm{I}(X_1 \curlyvee X_2 : Y) &< \mathrm{I}(X_1 : Y) + \mathrm{I}(X_2 : Y) - \mathrm{I}_\cap(X_1, X_2 : Y) \\ &= \mathrm{I}_\cup(X_1, X_2 : Y) \ . \end{aligned} \tag{3.9}$$

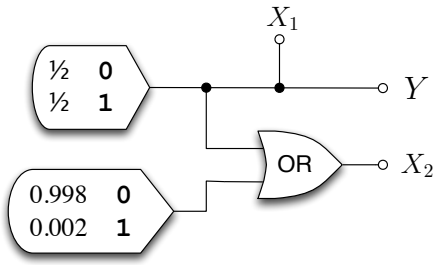
From eq. (3.9), we see negative synergy occurs when $\mathrm{I}_\cap$ is small, perhaps *too small*. Equivalently, negative synergy occurs when the joint r.v. conveys *less* about Y than the two r.v.'s X_1 and X_2 convey separately—mathematically, when $\mathrm{I}(X_1 \curlyvee X_2 : Y) < \mathrm{I}_\cup(X_1, X_2 : Y)$.[10] On the face of it this sounds strange. No usable structure in X_1 or X_2 "disappears" after they are combined by

[10] $\mathrm{I}_\cap$ and $\mathrm{I}_\cup$ are duals related by the inclusion–exclusion principle. For arbitrary n, this is $\mathrm{I}_\cup(X_1, \ldots, X_n : Y) = \sum_{\mathbf{S} \subseteq \{X_1, \ldots, X_n\}} (-1)^{|\mathbf{S}|+1} \mathrm{I}_\cap\Big(S_1, \ldots, S_{|\mathbf{S}|} : Y\Big)$.

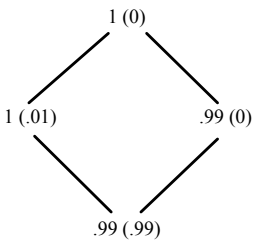
X_1	X_2	Y	
0	0	0	0.499
0	1	0	0.001
1	1	1	0.500

(a) $\Pr(x_1, x_2, y)$

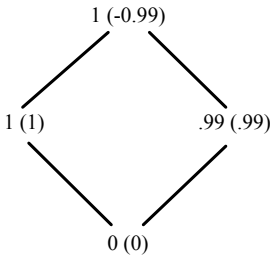
$$\mathrm{I}(X_1 \curlyvee X_2 : Y) = 1$$
$$\mathrm{I}(X_1 : Y) = 1$$
$$\mathrm{I}(X_2 : Y) = 0.99$$
$$\mathrm{I}_{\min}\big(\{X_1, X_2\} : Y\big) = 0.99$$
$$\mathrm{I}_{\curlywedge}(X_1, X_2 : Y) = 0$$



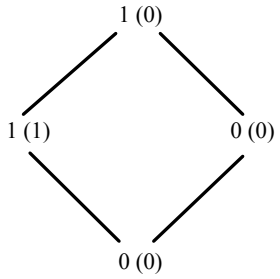
(b) circuit diagram



(c) $\mathrm{I}_{\min}$



(d) $\mathrm{I}_{\curlywedge}$



(e) $\mathrm{I}^0_{\curlywedge}$

Figure 3.5: Example ImperfectRdn. $\mathrm{I}_{\curlywedge}$ is blind to the noisy correlation between X_1 and X_2 and calculates zero redundant information. An ideal $\mathrm{I}_\cap$ measure would detect that all of the information X_2 specifies about Y is also specified by X_1 to calculate $\mathrm{I}_\cap(X_1, X_2 : Y) = 0.99$ bits.

$Z = X_1 \curlyvee X_2$. By the definition of $\curlyvee$, there are always functions f_1 and f_2 such that $X_1 \cong f_1(Z)$ and $X_2 \cong f_2(Z)$. Therefore, if your favorite $\mathrm{I}_\cap$ measure does not satisfy (**LP$_\mathbf{0}$**), it is likely too strict.

This means that, to our surprise, our measure $\mathrm{I}^0_\curlywedge$ does not account for the full zero-information overlap between $\mathrm{I}^0(X_1 : Y)$ and $\mathrm{I}^0(X_2 : Y)$. This is shown in example SUBTLE (Figure 3.6) where $\mathrm{I}^0_\curlywedge$ calculates a synergy of -0.252 bits. Defining a zero-error $\mathrm{I}_\cap$ that satisfies (**LP$_\mathbf{0}$**) is a matter of ongoing research.

3.7.1 Consequences of state-dependent (GP)

In [15] it's argued that $\mathrm{I}_{\min}$ upperbounds the ideal $\mathrm{I}_\cap$. Inspired by $\mathrm{I}_{\min}$ assuming state-dependent (**SR**) and (**M$_\mathbf{0}$**) to achieve a tighter upperbound on $\mathrm{I}_\cap$, we assume state-dependent (**GP**) to achieve a tighter lowerbound on $\mathrm{I}_\cap$ for $n = 2$. Our bound, denoted $\mathrm{I}_{\mathrm{smp}}$ for "sum minus pair", is defined as,

$$\mathrm{I}_{\mathrm{smp}}(X_1, X_2 : Y) \equiv \sum_{y \in Y} \Pr(y) \max\left[0, \mathrm{I}(X_1 : y) + \mathrm{I}(X_2 : y) - \mathrm{I}(X_1 \curlyvee X_2 : y)\right] , \tag{3.10}$$

where $\mathrm{I}(\bullet : y)$ is the same Kullback-Liebler divergence from eq. (3.1).

For example SUBTLE, the target $Y \cong X_1 \curlyvee X_2$, therefore per (**Id**), $\mathrm{I}_\cap(X_1, X_2 : Y) = \mathrm{I}(X_1 : X_2) = 0.252$ bits. However, given state-dependent (**GP**), applying $\mathrm{I}_{\mathrm{smp}}$ yields $\mathrm{I}_\cap(X_1, X_2 : Y) \geq 0.390$. Therefore, (**Id**) and state-dependent (**GP**) are incompatible. Secondly, given state-dependent (**GP**), example SUBTLE additionally illustrates a conjecture from [7] that the intersection information two predictors have about a target can exceed the mutual information between them, i.e., $\mathrm{I}_\cap(X_1, X_2 : Y) \not\leq \mathrm{I}(X_1 : X_2)$.

3.8 Conclusion and Path Forward

We've made incremental progress on several fronts towards the ideal Shannon $\mathrm{I}_\cap$.

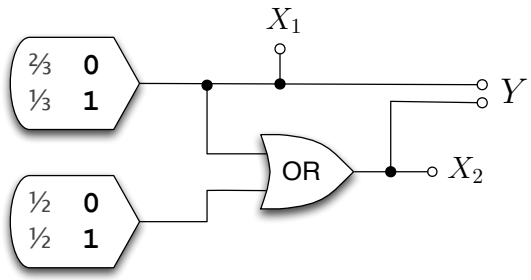
Desired Properties. We have tightened, expanded, and pruned the desired properties for $\mathrm{I}_\cap$. Particularly,

- (**LB**) is a non-contentious yet tighter lower-bound on $\mathrm{I}_\cap$ than (**GP**).
- Motivated by the natural equality $\mathrm{I}_\cap(X_1, \ldots, X_n : Y) = \mathrm{I}_\cap(X_1, \ldots, X_n, Y : Y)$, we introduce (**M$_\mathbf{1}$**) as a desired property.
- What was before an implicit assumption, we introduce (**Eq**) to better ground one's thinking.
- A separate chain-rule property is superfluous. Any desirable properties of conditional $\mathrm{I}_\cap$ are simply consequences of (**GP**) and (**TM**).

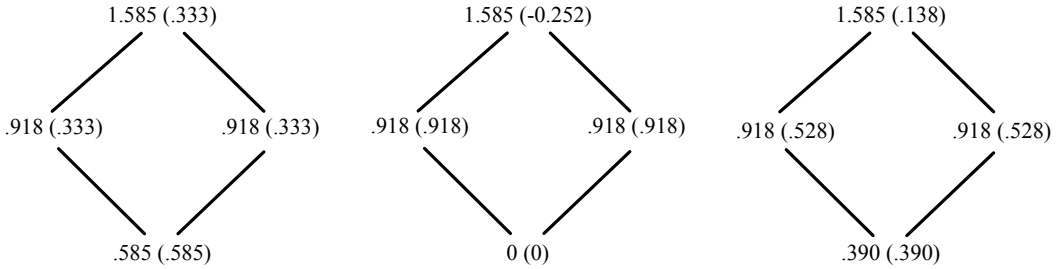
X_1 X_2	Y	
0 0	00	1/3
0 1	01	1/3
1 1	11	1/3

(a) $\Pr(x_1, x_2, y)$

$$\begin{aligned}
\mathrm{I}(X_1 \curlyvee X_2 : Y) &= 1.585 \\
\mathrm{I}(X_1 : Y) &= 0.918 \\
\mathrm{I}(X_2 : Y) &= 0.918 \\
\mathrm{I}(X_1 : X_2) &= 0.252 \\
\mathrm{I}_{\min}\big(\{X_1, X_2\} : Y\big) &= 0.585 \\
\mathrm{I}_{\curlywedge}(X_1, X_2 : Y) &= 0.0 \\
\mathrm{I}_{\mathrm{smp}}(X_1, X_2 : Y) &= 0.390
\end{aligned}$$



(b) circuit diagram



(c) $\mathrm{I}_{\min}$ (d) $\mathrm{I}_{\curlywedge}$ and $\mathrm{I}^0_{\curlywedge}$ (e) $\mathrm{I}_{\mathrm{smp}}$

Figure 3.6: Example SUBTLE. In this example both $\mathrm{I}_{\curlywedge}$ and $\mathrm{I}^0_{\curlywedge}$ calculate a synergy of -0.252 bits of synergy. What kind of redundancy must be captured for a nonnegative decomposition for this example?

A new measure. Based on the Gács-Körner common random variable, we introduced a new Shannon $\mathrm{I}_\cap$ measure. Our measure, $\mathrm{I}_\curlywedge$, is theoretically principled and the first to satisfy (**TM**).

How to improve. We identified where $\mathrm{I}_\curlywedge$ fails; it does not detect "imperfect" correlations between X_1 and X_2. One next step is to develop a less stringent $\mathrm{I}_\cap$ measure that satisfies ($\mathbf{LP_0}$) for simple nondeterministic examples like ImperfectRdn while still satisfying (**TM**).

To our surprise, example Subtle shows that $\mathrm{I}^0_\curlywedge$ does not satisfy ($\mathbf{LP_0}$)! This suggests that $\mathrm{I}^0_\curlywedge$ is too strict—what kind of zero-error informational overlap is $\mathrm{I}^0_\curlywedge$ not capturing? A separate next step is to formalize what exactly is required for a zero-error $\mathrm{I}_\cap$ to satisfy ($\mathbf{LP_0}$). From Subtle we can likewise see that for zero-error information, ($\mathbf{LP_0}$) is incompatible with (**Id**).

Finally, we showed that state-dependent (**GP**), a seemingly reasonable property, is incompatible with (**Id**) and moreover entails that $\mathrm{I}_\cap(X_1, X_2 : Y)$ can exceed $\mathrm{I}(X_1 : X_2)$.

Appendix

3.A Algorithm for Computing Common Random Variable

Given n random variables $X_1, \ldots, X_n$, the common random variable $X_1 \curlywedge \cdots \curlywedge X_n$ is computed by steps 1–3 in Appendix 3.B.

3.B Algorithm for Computing $\mathrm{I}_{\curlywedge}$

1. For each X_i for $i = 1, \ldots, n$, take its states x_i and place them as nodes on a graph. At the end of this process there will be $\sum_{i=1}^{n} |X_i|$ nodes on the graph.

2. For each pair of RVs X_i, X_j $(i \neq j)$, draw an undirected edge connecting nodes x_i and x_j if $\Pr\big(x_i, x_j\big) > 0$. At the end of this process the undirected graph will consist of k connected components $1 \leq k \leq \min_i |X_i|$. Denote these k disjoint components as $\mathbf{c_1}, \ldots, \mathbf{c_k}$.

3. Each connected component of the graph constitutes a distinct state of the common random variable Q, i.e., $|Q| = k$. Denote the states of the common random variable Q by $q_1, \ldots, q_k$.

4. Construct the joint probability distribution $\Pr(Q, Y)$ as follows. For every state $(q_i, y) \in Q \times Y$, the joint probability is created by summing over the entries of $\Pr(x_1, \ldots, x_n, y)$ in component i. More precisely,

$$\Pr(Q = q_i, Y = y) = \sum_{x_1, \ldots, x_n} \Pr(x_1, \ldots, x_n, y) \qquad \text{if } \{x_1, \ldots, x_n\} \subseteq \mathbf{c_i} \ .$$

5. Using $\Pr(Q, Y)$, compute $\mathrm{I}_{\curlywedge}(X_1, \ldots, X_n : Y)$ simply by computing the Shannon mutual information between Q and Y, i.e., $\mathrm{I}(Q : Y) = \mathrm{D}_{\mathrm{KL}}\left[\Pr(Q, Y) \middle\| \Pr(Q)\Pr(Y)\right]$.

3.C Lemmas and Proofs

3.C.1 Lemmas on Desired Properties

Lemma 3.C.1. *If* (**LB**) *holds, then* $\mathrm{I}_{\cap}(X_1, \ldots, X_n : Y) \geq \mathcal{I}\left(X_1 \curlywedge \cdots \curlywedge X_n : Y\right)$.

Proof. Assume that (**LB**) holds. By definition, $X_1 \curlywedge \cdots \curlywedge X_n \preceq X_i$ for $i = 1, \ldots, n$. So, by (**LB**), we immediately conclude that $\mathrm{I}_\cap(X_1, \ldots, X_n : Y) \geq \mathcal{I}(X_1 \curlywedge \cdots \curlywedge X_n : Y)$, which is the desired result. □

For the converse, we need the following assumption:

(**IM**) If $X_1 \preceq X_2$, then $\mathcal{I}(X_1 : Y) \leq \mathcal{I}(X_2 : Y)$.

Lemma 3.C.2. *Suppose that* (**IM**) *holds, and that* $\mathrm{I}_\cap(X_1, \ldots, X_n : Y) \geq \mathcal{I}(X_1 \curlywedge \cdots \curlywedge X_n : Y)$. *Then* (**LB**) *holds.*

Proof. Assume that $\mathrm{I}_\cap(X_1, \ldots, X_n : Y) \geq \mathcal{I}(X_1 \curlywedge \cdots \curlywedge X_n : Y)$. Let $Q \preceq X_i$ for $i = 1, \ldots, n$. Because $X_1 \curlywedge \cdots \curlywedge X_n$ is the largest (informationally richest) random variable that is informationally poorer than X_i for $i = 1, \ldots, n$, it follows that $Q \preceq X_1 \curlywedge \cdots \curlywedge X_n$. Therefore, by (**IM**), $\mathcal{I}(X_1 \curlywedge \cdots \curlywedge X_n : Y) \geq \mathcal{I}(Q : Y)$. Hence, $\mathrm{I}_\cap(X_1, \ldots, X_n : Y) \geq \mathcal{I}(Q : Y)$ also, which completes the proof. □

Remark: Assumption (**IM**) is satisfied by zero-error information and Shannon mutual information.

Lemma 3.C.3. *Given* $\mathrm{I}_\cap$, $X_1, \ldots, X_n$, Y, *and* Z, *consider the conditional intersection information*

$$\mathrm{I}_\cap(X_1, \ldots, X_n : Z|Y) = \mathrm{I}_\cap(X_1, \ldots, X_n : Y \curlyvee Z) - \mathrm{I}_\cap(X_1, \ldots, X_n : Y)\,.$$

Suppose that (**GP**), (**Eq**), *and* (**TM**) *hold. Then, the following properties hold:*

- $\mathrm{I}_\cap(X_1, \ldots, X_n : Z|Y) \geq 0$.
- $\mathrm{I}_\cap(X_1, \ldots, X_n : Z|Y) = \mathrm{I}_\cap(X_1, \ldots, X_n : Z)$ *if* Y *is a constant.*

Proof. We have $Y \preceq Y \curlyvee Z$. Therefore, by (**TM**), it immediately follows that $\mathrm{I}_\cap(X_1, \ldots, X_n : Z|Y) \geq 0$.

Next, suppose that Y is a constant. Then $Y \preceq Z$, and hence $Y \curlyvee Z \cong Z$. By (**Eq**), $\mathrm{I}_\cap(X_1, \ldots, X_n : Y \curlyvee Z) = \mathrm{I}_\cap(X_1, \ldots, X_n : Z)$. Moreover, by (**GP**), $\mathrm{I}_\cap(X_1, \ldots, X_n : Y) = 0$. Thus, $\mathrm{I}_\cap(X_1, \ldots, X_n : Z|Y) = \mathrm{I}_\cap(X_1, \ldots, X_n : Z)$ as desired. □

3.C.2 Properties of $\mathrm{I}^0_\curlywedge$

Lemma 3.C.4. *The measure of intersection information* $\mathrm{I}^0_\curlywedge(X_1, \ldots, X_n : Y)$ *satisfies* (**GP**), (**Eq**), (**TM**), (**M$_0$**), *and* (**S$_0$**), *but not* (**LP$_0$**).

Proof. (**GP**): The inequality $\mathrm{I}^0_{\curlywedge}(X_1, \ldots, X_n : Y) \geq 0$ follows immediately from the identity $\mathrm{I}^0_{\curlywedge}(X_1, \ldots, X_n : Y) = \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y)$ and the nonnegativity of $\mathrm{H}(\cdot)$. Next, if Y is a constant, then by the generalized absorption law, $X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y \cong Y$. Thus, by the invariance of $\mathrm{H}(\cdot)$ (Lemma 3.3.1(a)), $\mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y) = \mathrm{H}(Y) = 0$.

(**Eq**): Consider $X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y$. The equivalence class (with respect to $\cong$) in which this random variable resides is closed under substitution of X_i (for $i = 1, \ldots, n$) or Y by an informationally equivalent random variable. Hence, because $\mathrm{I}^0_{\curlywedge}(X_1, \ldots, X_n : Y) = \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y)$ and $\mathrm{H}(\cdot)$ is invariant over the equivalence class of random variables that are informationally equivalent to $X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y$ (by Lemma 3.3.1(a)), the desired result holds.

(**TM**): Suppose that $Y \preceq Z$. Then, $X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y \preceq X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Z$. Then, we have

$$\begin{aligned}
\mathrm{I}^0_{\curlywedge}(X_1, \ldots, X_n : Y) &= \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y) \\
&\leq \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Z) \quad \text{by monotonicity of } \mathrm{H}(\cdot) \text{ (Lemma 3.3.1(b))} \\
&= \mathrm{I}^0_{\curlywedge}(X_1, \ldots, X_n : Z)\,,
\end{aligned}$$

as desired.

(**$\mathbf{M_0}$**): By the generalized absorption law, $X_1 \curlywedge \cdots \curlywedge X_n \curlywedge W \curlywedge Y \preceq X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y$. Hence,

$$\begin{aligned}
\mathrm{I}^0_{\curlywedge}(X_1, \ldots, X_n, W : Y) &= \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge W \curlywedge Y) \\
&\leq \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y) \quad \text{by monotonicity of } \mathrm{H}(\cdot) \text{ (Lemma 3.3.1(b))} \\
&= \mathrm{I}^0_{\curlywedge}(X_1, \ldots, X_n : Y)\,,
\end{aligned}$$

as desired.

Next, suppose that there exists $Z \in \{X_1, \ldots, X_n\}$ such that $Z \preceq W$. Then, by the generalized absorption law, $X_1 \curlywedge \cdots \curlywedge X_n \curlywedge W \curlywedge Y \cong X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y$. Hence,

$$\begin{aligned}
\mathrm{I}^0_{\curlywedge}(X_1, \ldots, X_n, W : Y) &= \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge W \curlywedge Y) \\
&= \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y) \quad \text{by invariance of } \mathrm{H}(\cdot) \text{ (Lemma 3.3.1(a))} \\
&= \mathrm{I}^0_{\curlywedge}(X_1, \ldots, X_n : y)\,,
\end{aligned}$$

as desired.

(**$\mathbf{S_0}$**): By the commutativity law, $X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y$ is invariant (with respect to $\cong$) under reordering of $X_1, \ldots, X_n$. Hence, the desired result follows immediately from the identity $\mathrm{I}^0_{\curlywedge}(X_1, \ldots, X_n : Y) = \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y)$ and the invariance of $\mathrm{H}(\cdot)$ (Lemma 3.3.1(a)).

(**LP$_\mathbf{0}$**): For $\mathrm{I}^0_\curlywedge$, (**LP$_\mathbf{0}$**) relative to zero-error information can be written as

$$\mathrm{H}(X_1 \curlywedge X_2 \curlywedge Y) \geq \mathrm{H}(X_1 \curlywedge Y) + \mathrm{H}(X_2 \curlywedge Y) - \mathrm{H}\big((X_1 \curlyvee X_2) \curlywedge Y\big)\,. \tag{3.11}$$

However, this inequality does not hold in general. To see this, suppose that it does hold for arbitrary X_1, X_2, and Y. Note that $(X_1 \curlyvee X_2) \curlywedge Y \preceq Y$, which implies that $\mathrm{H}\big((X_1 \curlyvee X_2) \curlywedge Y\big) \leq \mathrm{H}(Y)$ (by monotonicity of $\mathrm{H}(\cdot)$). Hence, the inequality (3.11) implies that

$$\mathrm{H}(X_1 \curlywedge X_2 \curlywedge Y) \geq \mathrm{H}(X_1 \curlywedge Y) + \mathrm{H}(X_2 \curlywedge Y) - \mathrm{H}(Y)\,.$$

Rewriting this, we get

$$\mathrm{H}(X_1 \curlywedge Y) + \mathrm{H}(Y \curlywedge X_2) \leq \mathrm{H}(X_1 \curlywedge Y \curlywedge X_2) + \mathrm{H}(Y)\,.$$

But this is the supermodularity law for common information, which is known to be false in general; see [22], Section 5.4.

□

Lemma 3.C.5. *With respect to zero-error information, the measure of intersection information* $\mathrm{I}^0_\curlywedge(X_1, \ldots, X_n : Y)$ *satisfies* (**LB**)*,* (**SR**)*, and* (**Id**)*.*

Proof. (**LB**): Suppose that $Q \preceq X_i$ for $i = 1, \ldots, n$. Because $X_1 \curlywedge \cdots \curlywedge X_n$ is the largest (informationally richest) random variable that is informationally poorer than X_i for $i = 1, \ldots, n$, it follows that $Q \preceq X_1 \curlywedge \cdots \curlywedge X_n$. This implies that $X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y \succeq Q \curlywedge Y$. Therefore,

$$\begin{aligned}
\mathrm{I}^0_\curlywedge(X_1, \ldots, X_n : Y) &= \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y) \\
&\geq \mathrm{H}(Q \curlywedge Y) \quad \text{by monotonicity of } \mathrm{H}(\cdot) \text{ (Lemma 3.3.1(b))} \\
&= \mathrm{I}^0(Q : Y)\,,
\end{aligned}$$

as desired.

(**SR**): We have $\mathrm{I}^0_\curlywedge(X_1 : Y) = \mathrm{H}(X_1 \curlywedge Y) = \mathrm{I}^0(X_1 : Y)$.

(**Id**): By the associative and absorption laws, we have $X \curlywedge Y \curlywedge (X \curlyvee Y) \cong X \curlywedge Y$. Thus,

$$\begin{aligned}
\mathrm{I}^0_\curlywedge(X, Y : X \curlyvee Y) &= \mathrm{H}\big(X \curlywedge Y \curlywedge (X \curlyvee Y)\big) \\
&= \mathrm{H}(X \curlywedge Y) \quad \text{by invariance of } \mathrm{H}(\cdot) \text{ (Lemma 3.3.1(a))} \\
&= \mathrm{I}^0(X : Y)\,,
\end{aligned}$$

as desired.

□

Lemma 3.C.6. *The measure of intersection information* $\mathrm{I}^0_{\curlywedge}(X_1, \ldots, X_n : Y)$ *satisfies* $(\mathbf{M_1})$ *and* $(\mathbf{S_1})$*, but not* $(\mathbf{LP_1})$*.*

Proof. $(\mathbf{M_1})$: The desired inequality is identical to $(\mathbf{M_0})$, so it remains to prove the sufficient condition for equality. Suppose that there exists $Z \in \{X_1, \ldots, X_n, Y\}$ such that $Z \preceq W$. Then, by the generalized absorption law, $X_1 \curlywedge \cdots \curlywedge X_n \curlywedge W \curlywedge Y \cong X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Z$. Hence,

$$\begin{aligned}
\mathrm{I}^0_{\curlywedge}(X_1, \ldots, X_n, W : Y) &= \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge W \curlywedge Y) \\
&= \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Z) \quad \text{by invariance of } \mathrm{H}(\cdot) \text{ (Lemma 3.3.1(a))} \\
&= \mathrm{I}^0_{\curlywedge}(X_1, \ldots, X_n : Z)\,,
\end{aligned}$$

as desired.

$(\mathbf{S_1})$: By the commutativity law, $X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y$ is invariant (with respect to $\cong$) under reordering of $X_1, \ldots, X_n, Y$. Hence, the desired result follows immediately from the identity $\mathrm{I}^0_{\curlywedge}(X_1, \ldots, X_n : Y) = \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y)$ and the invariance of $\mathrm{H}(\cdot)$ (Lemma 3.3.1(a)).

$(\mathbf{LP_1})$: This follows from not satisfying $(\mathbf{LP_0})$.

□

3.C.3 Properties of $\mathrm{I}_{\curlywedge}$

Lemma 3.C.7. *The measure of intersection information* $\mathrm{I}_{\curlywedge}(X_1, \ldots, X_n : Y)$ *satisfies* $(\mathbf{GP})$*,* $(\mathbf{Eq})$*,* $(\mathbf{TM})$*,* $(\mathbf{M_0})$*, and* $(\mathbf{S_0})$*, but not* $(\mathbf{LP_0})$*.*

Proof. $(\mathbf{GP})$: The inequality $\mathrm{I}_{\curlywedge}(X_1, \ldots, X_n : Y) \geq 0$ follows immediately from the identity $\mathrm{I}_{\curlywedge}(X_1, \ldots, X_n : Y) = \mathrm{I}(X_1 \curlywedge \cdots \curlywedge X_n : Y)$ and the nonnegativity of mutual information. Next, suppose that Y is a constant. Then $\mathrm{H}(Y) = 0$. Moreover, $Y \preceq X_1 \curlywedge \cdots \curlywedge X_n$ by definition of $\curlywedge$. Thus, by Lemma 3.3.1(c), $\mathrm{H}\big(Y | X_1 \curlywedge \cdots \curlywedge X_n\big) = 0$, and

$$\begin{aligned}
\mathrm{I}_{\curlywedge}(X_1, \ldots, X_n : Y) &= \mathrm{I}(X_1 \curlywedge \cdots \curlywedge X_n : Y) \\
&= \mathrm{I}(Y : X_1 \curlywedge \cdots \curlywedge X_n) \\
&= \mathrm{H}(Y) - \mathrm{H}\big(Y | X_1 \curlywedge \cdots \curlywedge X_n\big) \\
&= 0.
\end{aligned}$$

$(\mathbf{Eq})$: Consider $X_1 \curlywedge \cdots \curlywedge X_n$. The equivalence class (with respect to $\cong$) in which this random variable resides is closed under substitution of X_i (for $i = 1, \ldots, n$) or Y by an informationally

equivalent random variable. Hence, because

$$\begin{aligned}\mathrm{I}_{\curlywedge}(X_1,\ldots,X_n:Y) &= \mathrm{H}(Y) - \mathrm{H}\big(Y|X_1 \curlywedge \cdots \curlywedge X_n\big) \\ &= \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n) - \mathrm{H}\big(X_1 \curlywedge \cdots \curlywedge X_n|Y\big)\,,\end{aligned}$$

by Lemma 3.3.1(a), the desired result holds.

(**TM**): Suppose that $Y \preceq Z$. For simplicity, let $Q = X_1 \curlywedge \cdots \curlywedge X_n$. Then,

$$\begin{aligned}\mathrm{I}_{\curlywedge}(X_1,\ldots,X_n:Y) &= \mathrm{H}(Q) - \mathrm{H}\big(Q|Y\big) \\ &\leq \mathrm{H}(Q) - \mathrm{H}\big(Q|Z\big) \quad \text{by Lemma 3.3.1(b)} \\ &= \mathrm{I}_{\curlywedge}(X_1,\ldots,X_n:Z)\,,\end{aligned}$$

as desired.

(**$\mathbf{M_0}$**): By definition of $\curlywedge$, we have $X_1 \curlywedge \cdots \curlywedge X_n \curlywedge W \preceq X_1 \curlywedge \cdots \curlywedge X_n$. Hence,

$$\begin{aligned}\mathrm{I}_{\curlywedge}(X_1,\ldots,X_n,W:Y) &= \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge W) - \mathrm{H}\big(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge W|Y\big) \\ &\leq \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n) - \mathrm{H}\big(X_1 \curlywedge \cdots \curlywedge X_n|Y\big) \quad \text{by Lemma 3.3.1(b)} \\ &= \mathrm{I}_{\curlywedge}(X_1,\ldots,X_n:Y)\,,\end{aligned}$$

as desired.

Next, suppose that there exists $Z \in \{X_1,\ldots,X_n\}$ such that $Z \preceq W$. Then, by the algebraic laws of $\curlywedge$, we have $X_1 \curlywedge \cdots \curlywedge X_n \curlywedge W \cong X_1 \curlywedge \cdots \curlywedge X_n$. Hence,

$$\begin{aligned}\mathrm{I}_{\curlywedge}(X_1,\ldots,X_n,W:Y) &= \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge W) - \mathrm{H}\big(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge W|Y\big) \\ &= \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n) - \mathrm{H}\big(X_1 \curlywedge \cdots \curlywedge X_n|Y\big) \quad \text{by Lemma 3.3.1(a)} \\ &= \mathrm{I}_{\curlywedge}(X_1,\ldots,X_n:Y)\,,\end{aligned}$$

as desired.

(**$\mathbf{S_0}$**): By the commutativity law, $X_1 \curlywedge \cdots \curlywedge X_n$ is invariant (with respect to $\cong$) under reordering of $X_1,\ldots,X_n$. Hence, the desired result follows immediately from the identity $\mathrm{I}_{\curlywedge}(X_1,\ldots,X_n:Y) = \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n) - \mathrm{H}\big(X_1 \curlywedge \cdots \curlywedge X_n|Y\big)$ and Lemma 3.3.1(a).

(**$\mathbf{LP_0}$**): A counterexample is provided by ImperfectRdn (Figure 3.5).

□

Lemma 3.C.8. *With respect to mutual information, the measure of intersection information* $\mathrm{I}_{\curlywedge}(X_1,\ldots,X_n:Y)$ *satisfies* (**LB**) *and* (**SR**)*, but not* (**Id**)*.*

Proof. (**LB**): Suppose that $Q \preceq X_i$ for $i = 1, \ldots, n$. Because $X_1 \curlywedge \cdots \curlywedge X_n$ is the largest (informationally richest) random variable that is informationally poorer than X_i for $i = 1, \ldots, n$, it follows that $Q \preceq X_1 \curlywedge \cdots \curlywedge X_n$. Therefore,

$$\begin{aligned} \mathrm{I}_\curlywedge(X_1, \ldots, X_n : Y) &= \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n) - \mathrm{H}\big(X_1 \curlywedge \cdots \curlywedge X_n | Y\big) \\ &\geq \mathrm{H}(Q) - \mathrm{H}\big(Q|Y\big) \quad \text{by Lemma 3.3.1(b)} \\ &= \mathrm{I}(Q{:}Y)\,, \end{aligned}$$

as desired.

(**SR**): By definition, $\mathrm{I}_\curlywedge(X_1 : Y) = \mathrm{I}(X_1{:}Y)$.

(**Id**): We have $X \curlywedge Y \preceq X \curlyvee Y$ by definition of $\curlywedge$ and $\curlyvee$. Thus,

$$\begin{aligned} \mathrm{I}_\curlywedge(X, Y : X \curlyvee Y) &= \mathrm{I}(X \curlywedge Y{:}X \curlyvee Y) \\ &= \mathrm{H}(X \curlywedge Y) - \mathrm{H}\big(X \curlywedge Y | X \curlyvee Y\big) \\ &= \mathrm{H}(X \curlywedge Y) \quad \text{by Lemma 3.3.1(a)} \\ &= \mathrm{I}^0(X{:}Y) \\ &\neq \mathrm{I}(X{:}Y)\ . \end{aligned}$$

□

Lemma 3.C.9. *The measure of intersection information* $\mathrm{I}_\curlywedge(X_1, \ldots, X_n{:}Y)$ *does not satisfy* (**M₁**)*,* (**S₁**)*, and* (**LP₁**)*.*

Proof. (**M₁**): A counterexample is provided in IMPERFECTRDN (Figure 3.5), where $\mathrm{I}_\curlywedge(X_1 : Y) = 0.99$ bits, yet $\mathrm{I}_\curlywedge(X_1, Y : Y) = 0$ bits.

(**S₁**): A counterexample. We show $\mathrm{I}_\curlywedge(X, X{:}Y) \neq \mathrm{I}_\curlywedge(X, Y{:}X)$.

$$\begin{aligned} \mathrm{I}_\curlywedge(X, X{:}Y) - \mathrm{I}_\curlywedge(X, Y{:}X) &= \mathrm{I}(X{:}Y) - \mathrm{I}_\curlywedge(X, Y{:}X) \\ &= \mathrm{I}(X{:}Y) - \mathrm{I}(X \curlywedge Y{:}X) \\ &= \mathrm{I}(X{:}Y) - \mathrm{H}(X \curlywedge Y) - \mathrm{H}\big(X \curlywedge Y | X\big) \\ &= \mathrm{I}(X{:}Y) - \mathrm{H}(X \curlywedge Y) \\ &\neq 0\ . \end{aligned}$$

(**LP₁**): This follows from not satisfying (**LP₀**).

□

3.D Miscellaneous Results

Simplification of $\mathrm{I}^0_{\curlywedge}$

Lemma 3.D.1. *We have* $\mathrm{I}^0_{\curlywedge}(X_1,\ldots,X_n{:}Y) = \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y)$.

Proof. By definition,

$$\begin{aligned}\mathrm{I}^0_{\curlywedge}(X_1,\ldots,X_n{:}Y) \equiv & \max_{\Pr(Q|Y)} \mathrm{I}^0(Q{:}Y) \\ & \text{subject to } Q \preceq X_i \ \forall i \in \{1,\ldots,n\} \\ = & \max_{\Pr(Q|Y)} \mathrm{H}(Q \curlywedge Y) \\ & \text{subject to } Q \preceq X_i \ \forall i \in \{1,\ldots,n\}\end{aligned}$$

Let Q be an arbitrary random variable satisfying the constraint $Q \preceq X_i$ for $i = 1,\ldots,n$. Because $X_1 \curlywedge \cdots \curlywedge X_n$ is the largest random variable (in the sense of the partial order $\preceq$) that is informationally poorer than X_i for $i = 1,\ldots,n$, we have $Q \preceq X_1 \curlywedge \cdots \curlywedge X_n$. By the property of $\curlywedge$ pointed out before, we also have $Q \curlywedge Y \preceq X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y$. By Lemma 3.3.1(b), this implies that $\mathrm{H}(Q \curlywedge Y) \leq \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y)$. Therefore, $\mathrm{I}^0_{\curlywedge}(X_1,\ldots,X_n{:}Y) = \mathrm{H}(X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y)$. □

Simplification of $\mathrm{I}_{\curlywedge}$

Lemma 3.D.2. *We have* $\mathrm{I}_{\curlywedge}(X_1,\ldots,X_n{:}Y) = \mathrm{I}(X_1 \curlywedge \cdots \curlywedge X_n{:}Y)$.

Proof. By definition,

$$\begin{aligned}\mathrm{I}_{\curlywedge}(X_1,\ldots,X_n{:}Y) \equiv & \max_{\Pr(Q|Y)} \mathrm{I}(Q{:}Y) \\ & \text{subject to } Q \preceq X_i \ \forall i \in \{1,\ldots,n\} \\ = & \, \mathrm{H}(Y) - \min_{\Pr(Q|Y)} \mathrm{H}\big(Y|Q\big) \\ & \text{subject to } Q \preceq X_i \ \forall i \in \{1,\ldots,n\}\end{aligned}$$

Let Q be an arbitrary random variable satisfying the constraint $Q \preceq X_i$ for $i = 1,\ldots,n$. Because $X_1 \curlywedge \cdots \curlywedge X_n$ is the largest random variable (in the sense of the partial order $\preceq$) that is informationally poorer than X_i for $i = 1,\ldots,n$, we have $Q \preceq X_1 \curlywedge \cdots \curlywedge X_n$. By Lemma 3.3.1(b), this implies that $\mathrm{H}\big(Y|Q\big) \geq \mathrm{H}\big(Y|X_1 \curlywedge \cdots \curlywedge X_n \curlywedge Y\big)$. Therefore, $\mathrm{I}_{\curlywedge}(X_1,\ldots,X_n{:}Y) = \mathrm{I}(X_1 \curlywedge \cdots \curlywedge X_n{:}Y)$. □

Proof that $\mathrm{I}_{\curlywedge}(X_1,\ldots,X_n{:}Y) \leq \mathrm{I}_{\min}(X_1,\ldots,X_n : Y)$

Lemma 3.D.3. *We have* $\mathrm{I}_{\curlywedge}(X_1,\ldots,X_n{:}Y) \leq \mathrm{I}_{\min}(X_1,\ldots,X_n : Y)$

Proof. Starting from the definitions,

$$\begin{aligned} \mathrm{I}_{\curlywedge}(X_1, \ldots, X_n : Y) &\equiv \mathrm{I}(X_1 \curlywedge \cdots \curlywedge X_n : Y) \\ &= \sum_y \Pr(y)\, \mathrm{I}(X_1 \curlywedge \cdots \curlywedge X_n : y) \\ \mathrm{I}_{\min}\big(\{X_1, \ldots, X_n\} : Y\big) &\equiv \sum_y \Pr(y) \min_i \mathrm{I}(X_i : y) \ . \end{aligned}$$

For a particular state y, without loss of generality we define the minimizing predictor X_m by $X_m \equiv \operatorname{argmin}_{X_i} \mathrm{I}(X_i : y)$ and the common random variable $Q \equiv X_1 \curlywedge \cdots \curlywedge X_n$. It then remains to show that $\mathrm{I}(Q : y) \leq \mathrm{I}(X_m : y)$.

By definition of $\curlywedge$, we have $Q \preceq X_m$. Hence,

$$\begin{aligned} \mathrm{I}(X_m : y) &= \mathrm{H}(X_m) - \mathrm{H}\big(X_m | Y = y\big) \\ &\geq \mathrm{H}(Q) - \mathrm{H}\big(Q | Y = y\big) \quad \text{by Lemma 3.3.1(b)} \\ &= \mathrm{I}(Q : y) \ . \end{aligned}$$

□

State-dependent zero-error information

We define the state-dependent zero-error information, $\mathrm{I}^0(X : Y = y)$ as,

$$\mathrm{I}^0(X : Y = y) \equiv \log \frac{1}{\Pr(Q = q)} \ ,$$

where the random variable $Q \equiv X \curlywedge Y$ and $\Pr(Q = q)$ is the probability of the connected component containing state $y \in Y$. This entails that $\Pr(y) \leq \Pr(q) \leq 1$. Similar to the state-dependent information, $\mathbb{E}_Y\, \mathrm{I}^0(X : y) \ = \ \mathrm{I}^0(X : Y)$, where $\mathbb{E}_Y$ is the expectation value over Y.

Proof. We define two functions f and g:

- $f : y \to q$ s.t. $\Pr\big(q|y\big) = 1$ where $q \in Q$ and $y \in Y$.
- $g : q \to \{y_1, \ldots, y_k\}$ s.t. $\Pr\big(q|y_i\big) = 1$ where $q \in Q$ and $y \in Y$.

Now we have,

$$\mathbb{E}_Y\, \mathrm{I}^0(X : y) \equiv \sum_{y \in Y} \Pr(y) \log \frac{1}{\Pr\big(f(y)\big)} \ .$$

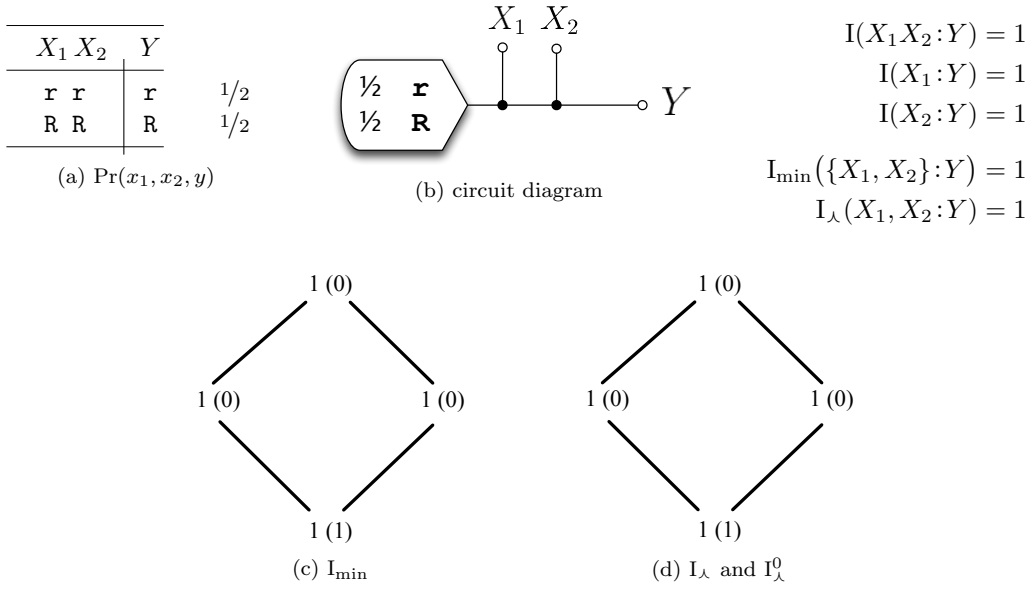
Since each y is associated with exactly one q, we can reindex the $\sum_{y \in Y}$. We then simplify to

achieve the result.

$$\begin{aligned}
\sum_{y\in Y} \Pr(y) \log \frac{1}{\Pr\bigl(f(y)\bigr)} &= \sum_{q\in Q} \sum_{y\in g(q)} \Pr(y) \log \frac{1}{\Pr\bigl(f(y)\bigr)} \\
&= \sum_{q\in Q} \sum_{y\in g(q)} \Pr(y) \log \frac{1}{\Pr(q)} = \sum_{q\in Q} \log \frac{1}{\Pr(q)} \sum_{y\in g(q)} \Pr(y) \\
&= \sum_{q\in Q} \log \frac{1}{\Pr(q)} \Pr(q) = \sum_{q\in Q} \Pr(q) \log \frac{1}{\Pr(q)} \\
&= \mathrm{H}(Q) = \mathrm{I}^0(X\!:\!Y) \ .
\end{aligned}$$

□

3.E Misc Figures



(a) $\Pr(x_1, x_2, y)$

(b) circuit diagram

(c) $\mathrm{I}_{\min}$

(d) $\mathrm{I}_\wedge$ and $\mathrm{I}^0_\wedge$

Figure 3.7: Example Rdn. In this example $\mathrm{I}_{\min}$ and $\mathrm{I}_\wedge$ reach the same answer yet diverge drastically for example ImperfectRdn.

Chapter 4

Irreducibility is Minimum Synergy among Parts

In this chapter we explore how a collective action can be "irreducible to the actions performed by its parts". First, we show that computing synergy among four common notions of "parts" gives rise to a spectrum of four distinct measures of irreducibility. Second, using Partial Information Decomposition[36], we introduce a nonnegative expression for each notion of irreducibility. Third, we delineate these four notions of irreducibility with exemplary binary circuits.

4.1 Introduction

Before we discussed computing synergy among random variables. Now we show that we can define broader notions of irreducibility by computing synergy among joint random variables. Therefore, a measure of synergy will allow us to quantify a myriad of notions of irreducibility. One pertinent application of quantifying irreducibility is finding the most useful granularity for analyzing a complex system in which interactions occur at multiple scales. Prior work [6, 19, 1, 36] has proposed measures of irreducibility, but there remains no consensus which measure is most valid.

4.1.1 Notation

In our treatment of irreducibility, the n agents are random variables $\{X_1, \ldots, X_n\}$, and the collective action the agents perform is predicting (having mutual information about) a single target random variable Y. We use the following notation throughout. Let

$\mathbf{X}$: The set of n elementary random variables (r.v.). $\mathbf{X} \equiv \{X_1, X_2, \ldots, X_n\}$. $n \geq 2$.

$X_{1\ldots n}$: The *whole*, the joint r.v. (cartesian product) of all n elements, $X_{1\ldots n} \equiv X_1 \curlyvee \cdots \curlyvee X_n$.

Y: The "target" random variable to be predicted.

$\mathcal{P}(\mathbf{X})$: The set of all parts (random variables) derivable from a nonempty, proper subset of $\mathbf{X}$. For a set of n elements there are 2^n-2 possible parts. Formally, $\mathcal{P}(\mathbf{X}) \equiv \Big\{S_1 \curlyvee \cdots \curlyvee S_{|\mathbf{S}|} : \mathbf{S} \subset \mathbf{X}, \mathbf{S} \neq \emptyset\Big\}$.

$\mathbf{P}$: A set of m parts $\mathbf{P} \equiv \{P_1, P_2, \ldots, P_m\}$, $2 \leq m \leq n$. Each part P_i is an element (random variable) in the set $\mathcal{P}(\mathbf{X})$. The joint random variable of all m parts is always informationally equivalent to $X_{1\ldots n}$, i.e., $P_1 \curlyvee \cdots \curlyvee P_m \cong X_{1\ldots n}$. Hereafter, the terms "part" and "component" are used interchangeably.

A_i: The i'th "Almost". An "Almost" is a part (joint random variable) only lacking the element X_i. $1 \leq i \leq n$. Formally, $A_i \equiv X_1 \curlyvee \cdots \curlyvee X_{i-1} \curlyvee X_{i+1} \curlyvee \cdots \curlyvee X_n$.

All capital letters are random variables. All bolded capital betters are sets of random variables.

4.2 Four common notions of irreducibility

Prior literature [11, 1, 6, 29] has intuitively conceptualized the irreducibility of the information a whole $X_{1\ldots n}$ conveys about Y in terms of how much information about Y is lost upon "breaking up" $X_{1\ldots n}$ into a set of parts $\mathbf{P}$. We express this intuition formally by computing the aggregate information $\mathbf{P}$ has about Y, and then subtracting it from the mutual information $\mathrm{I}(X_{1\ldots n}:Y)$. But what are the parts $\mathbf{P}$? The four most common choices are:

1. **The singleton elements**. We take the set of n elements, $\mathbf{X}$, compute the mutual information with Y when all n elements work separately, and then subtract it from $\mathrm{I}(X_{1\ldots n}:Y)$. Information beyond the Elements (IbE) is the weakest notion of irreducibility. In the PI-diagram[36] of $\mathrm{I}(X_{1\ldots n}:Y)$, IbE is the sum of all synergistic PI-regions.

2. **Any partition of (disjoint) parts**. We enumerate all possible partitions of set $\mathbf{X}$. Formally, a partition $\mathbf{P}$ is any set of parts $\{P_1, \ldots, P_m\}$ such that, $P_i \curlywedge P_j \prec X_k$ where $i, j \in \{1, \ldots, m\}$, $i \neq j$, and $k \in \{1, \ldots, n\}$. For each partition, we compute the mutual information with Y when its m parts work separately. We then take the maximum information over all partitions and subtract it from $\mathrm{I}(X_{1\ldots n}:Y)$. Information beyond the Disjoint Parts (IbDp) quantifies $\mathrm{I}(X_{1\ldots n}:Y)$'s irreducibility to information conveyed by disjoint parts.

3. **Any two parts**. We enumerate all "part-pairs" of set $\mathbf{X}$. Formally, a part-pair $\mathbf{P}$ is any set of exactly two elements in $\mathcal{P}(\mathbf{X})$. For each part-pair, we compute the mutual information with Y when the parts work separately. We then take the maximum mutual information over all part-pairs and subtract it from $\mathrm{I}(X_{1\ldots n}:Y)$. Information beyond the Two Parts ($\mathsf{Ib2p}$) quantifies $\mathrm{I}(X_{1\ldots n}:Y)$'s irreducibility to information conveyed by any pair of parts.

4. **All possible parts**. We take the set of all possible parts of set $\mathbf{X}$, $\mathcal{P}(\mathbf{X})$, and compute the information about Y conveyed when all $2^n - 2$ parts work separately and subtract it from $\mathrm{I}(X_{1\ldots n}:Y)$. Information beyond All Parts (IbAp) is the strongest notion of irreducibility. In the PI-diagram of $\mathrm{I}(X_{1\ldots n}:Y)$, IbAp is the value of PI-region $\{1\ldots n\}$.

4.3 Quantifying the four notions of irreducibility

To calculate the information in the whole beyond its elements, the first thing that comes to mind is to take the whole and subtract the sum over the elements, i.e., $\mathrm{I}(X_{1\ldots n}:Y) - \sum_{i=1}^{n} \mathrm{I}(X_i:Y)$. However, the sum *double-counts* when over multiple elements convey the same information about Y. To avoid double-counting the same information, we need to change the sum to "union". Whereas summing adds duplicate information multiple times, unioning adds duplicate information only once. This guiding intuition of "whole minus union" leads to the definition of irreducibility as the information conveyed by the whole minus the "union information" over its parts.

We provide expressions for IbE, IbDp, $\mathsf{Ib2p}$, and IbAp for arbitrary n. All four equations are the information conveyed by the whole, $\mathrm{I}(X_{1\ldots n}:Y)$, minus the maximum union information about Y over some parts $\mathbf{P}$, $\mathrm{I}_\cup(P_1,\ldots,P_m:Y)$. In PID, $\mathrm{I}_\cup$ is the dual to $\mathrm{I}_\cap$; they are related by the inclusion–exclusion principle. Thus if we only have a $\mathrm{I}_\cap$ measure we can always express the $\mathrm{I}_\cup$ by,

$$\mathrm{I}_\cup(P_1,\ldots,P_m:Y) = \sum_{\mathbf{S}\subseteq\{P_1,\ldots,P_m\}} (-1)^{|\mathbf{S}|+1}\, \mathrm{I}_\cap\Big(S_1,\ldots,S_{|\mathbf{S}|}:Y\Big) \ .$$

There are currently several candidate definitions of the union information[15, 14], but for our measures to work all that is required is that the $\mathrm{I}_\cup$ measure satisfy:

(**GP**) Global Positivity: $\mathrm{I}_\cup(P_1,\ldots,P_m:Y) \geq 0$, and $\mathrm{I}_\cup(P_1,\ldots,P_m:Y) = 0$ if Y is a constant.

(**Eq**) Equivalence-Class Invariance: $\mathrm{I}_\cup(P_1,\ldots,P_m:Y)$ is invariant under substitution of P_i (for any $i = 1,\ldots,m$) or Y by an informationally equivalent random variable.

($\mathbf{M_0}$) Weak Monotonicity: $\mathrm{I}_\cup(P_1,\ldots,P_m,W:Y) \geq \mathrm{I}_\cup(P_1,\ldots,P_m:Y)$ with equality if there exists $P_i \in \{P_1,\ldots,P_m\}$ such that $W \preceq P_i$.

($\mathbf{S_0}$) Weak Symmetry: $\mathrm{I}_\cup(P_1,\ldots,P_m:Y)$ is invariant under reordering of $P_1,\ldots,P_m$.

(**SR**) Self-Redundancy: $\mathrm{I}_\cup(P_1:Y) = \mathrm{I}(P_1:Y)$. The union information a single part P_1 conveys about the target Y is equal to the mutual information between P_1 and the target.

(**UB**) Upperbound: $\mathrm{I}_\cup(P_1,\ldots,P_m:Y) \leq \mathrm{I}(P_1 \curlyvee \cdots \curlyvee P_m:Y)$. In this particular case, the joint r.v. $P_1 \curlyvee \cdots \curlyvee P_m \cong X_{1\ldots n}$, so this equates to $\mathrm{I}_\cup(P_1,\ldots,P_m:Y) \leq \mathrm{I}(X_{1\ldots n}:Y)$.

4.3.1 Information beyond the Elements

Information beyond the Elements, $\mathsf{IbE}(\mathbf{X} : Y)$ quantifies how much information in $\mathrm{I}(X_{1\ldots n} : Y)$ isn't conveyed by any element X_i for $i \in \{1, \ldots, n\}$. The Information beyond the Elements is,

$$\mathsf{IbE}(\mathbf{X} : Y) \equiv \mathrm{I}(X_{1\ldots n} : Y) - \mathrm{I}_{\cup}(X_1, \ldots, X_n : Y) \ . \tag{4.1}$$

Information beyond the Elements, or *synergistic mutual information*[15], quantifies the amount of information in $\mathrm{I}(X_{1\ldots n} : Y)$ that *only coalitions* of elements convey.

4.3.2 Information beyond Disjoint Parts: IbDp

Information beyond Disjoint Parts, $\mathsf{IbDp}(\mathbf{X} : Y)$, quantifies how much information in $\mathrm{I}(X_{1\ldots n} : Y)$ isn't conveyed by any partition of set $\mathbf{X}$. Like IbE, IbDp is the total information minus the "union information" over components. Unlike IbE, the components are not the n elements but the parts of a partition. Some algebra proves that the partition with the maximum mutual information will always be a bipartition; thus we can safely restrict the maximization to bipartitions.[1] Therefore to quantify $\mathrm{I}(X_{1\ldots n} : Y)$'s irreducibility to disjoint parts, we maximize over all $2^{n-1} - 1$ bipartitions of set $\mathbf{X}$. Altogether, the Information beyond Disjoint Parts is,

$$\begin{aligned} \mathsf{IbDp}(\mathbf{X} : Y) &\equiv \mathrm{I}(X_{1\ldots n} : Y) - \max_{\substack{P_1 \in \mathcal{P}(\mathbf{X}) \\ \vdots \\ P_m \in \mathcal{P}(\mathbf{X}) \\ P_i \not\asymp P_j \prec X_k, \ \forall i \neq j \ \ k \in \{1,\ldots,n\}}} \mathrm{I}_{\cup}(P_1, \ldots, P_m : Y) && (4.2) \\ &= \mathrm{I}(X_{1\ldots n} : Y) - \max_{S \in \mathcal{P}(\mathbf{X})} \mathrm{I}_{\cup}\big(S, \mathbf{X} \setminus S : Y\big) \ . && (4.3) \end{aligned}$$

4.3.3 Information beyond Two Parts: Ib2p

Information beyond Two Parts, $\mathsf{Ib2p}(\mathbf{X} : Y)$, quantifies how much information in $\mathrm{I}(X_{1\ldots n} : Y)$ isn't conveyed by any pair of parts. Like IbDp, Ib2p subtracts the maximum union information over two parts. Unlike IbDp, the two parts aren't disjoint. Some algebra proves that the part-pair conveying the most information about Y will always be a pair of "Almosts".[2] Thus, we can safely restrict the maximization over all pairs of Almosts, and we maximize over the $\binom{n}{2} = \frac{n(n-1)}{2}$ pairs of Almosts. Altogether, the Information beyond Two Parts is,

$$\begin{aligned} \mathsf{Ib2p}(X_1, \ldots, X_n : Y) &\equiv \mathrm{I}(X_{1\ldots n} : Y) - \max_{\substack{P_1 \in \mathcal{P}(\mathbf{X}) \\ P_2 \in \mathcal{P}(\mathbf{X})}} \mathrm{I}_{\cup}(P_1, P_2 : Y) && (4.4) \\ &= \mathrm{I}(X_{1\ldots n} : Y) - \max_{\substack{i,j \in \{1,\ldots,n\} \\ i \neq j}} \mathrm{I}_{\cup}\big(A_i, A_j : Y\big) \ . && (4.5) \end{aligned}$$

[1] See Appendix 4.B.1 for a proof.
[2] See Appendix 4.B.2 for a proof.

4.3.4 Information beyond All Parts: IbAp

Information beyond All Parts, $\mathsf{IbAp}(\mathbf{X} : Y)$, quantifies how much information in $\mathrm{I}(X_{1\ldots n}{:}Y)$ isn't conveyed by any part. Like $\mathsf{Ib2p}$, IbAp subtracts the union information over overlapping parts. Unlike $\mathsf{Ib2p}$, the union is not over two parts, but all possible parts. Some algebra proves that the entirety of the information conveyed by all $2^n - 2$ parts working separately is equally conveyed by the n Almosts working separately.[3] Thus we can safely contract the union information to the n Almosts. Altogether, the Information beyond All Parts is,

$$\begin{aligned}\mathsf{IbAp}\left(X_1, \ldots, X_n : Y\right) &\equiv \mathrm{I}(X_{1\ldots n}{:}Y) - \mathrm{I}_\cup\big(\mathcal{P}(\mathbf{X}){:}Y\big) \\ &= \mathrm{I}(X_{1\ldots n}{:}Y) - \mathrm{I}_\cup(A_1, A_2, \ldots, A_n{:}Y) \ .\end{aligned} \tag{4.6}$$

Whereas Information beyond the Elements quantifies the amount of information in $\mathrm{I}(X_{1\ldots n}{:}Y)$ only conveyed by coalitions, Information beyond All Parts, or *holistic mutual information*, quantifies the amount of information in $\mathrm{I}(X_{1\ldots n}{:}Y)$ only conveyed by the whole.

By properties (**GP**) and (**UB**), our four measures are nonnegative and bounded by $\mathrm{I}(X_{1\ldots n}{:}Y)$. Finally, each succeeding of notion of components is a generalization of the prior. This successive generality gives rise to the handy inequality:

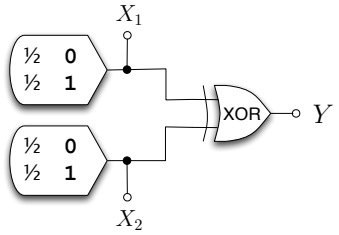
$$\mathsf{IbAp}(\mathbf{X} : Y) \leq \mathsf{Ib2p}(\mathbf{X} : Y) \leq \mathsf{IbDp}(\mathbf{X} : Y) \leq \mathsf{IbE}(\mathbf{X} : Y) \ . \tag{4.7}$$

4.4 Exemplary Binary Circuits

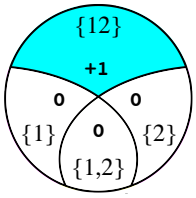
For $n = 2$, all four notions of irreducibility are equivalent; each one is simply the value of PI-region $\{12\}$ (see subfigures 4.2a–d). The canonical example of irreducibility for $n = 2$ is example Xor (Figure 4.1). In Xor, the irreducibility of X_1 and X_2 specifying Y is analogous to irreducibility of hydrogen and oxygen extinguishing fire. The whole X_1X_2 fully specifies Y, $\mathrm{I}(X_1X_2{:}Y) = \mathrm{H}(Y) = 1$ bit, but X_1 and X_2 separately convey nothing about Y, $\mathrm{I}(X_1{:}Y) = \mathrm{I}(X_2{:}Y) = 0$ bits.

X_1	X_2	Y	
0	0	0	1/4
0	1	1	1/4
1	0	1	1/4
1	1	0	1/4

(a) $\Pr(x_1, x_2, y)$



(b) circuit diagram



(c) PI-diagram

Figure 4.1: Example Xor. X_1X_2 irreducibly specifies Y. $\mathrm{I}(X_1X_2{:}Y) = \mathrm{H}(Y) = 1$ bit.

[3] See Appendix 4.B.3 for a proof.

For $n > 2$, the four notions of irreducibility diverge; subfigures 4.2e–j depicts IbE, IbAp, IbDp, and Ib2p when $n = 3$. We provide exemplary binary circuits delineating each measure. Every circuit has $n = 3$ elements, meaning $\mathbf{X} = \{X_1, X_2, X_3\}$, and build atop example Xor.

4.4.1 XorUnique: Irreducible to elements, yet reducible to a partition

To concretize how a collective action could be irreducible to elements, yet still reducible to a partition, consider a hypothetical set of agents $\{X_1, X_2, \ldots, X_{100}\}$ where the first 99 agents cooperate to specify Y, but agent X_{100} doesn't cooperate with the joint random variable $X_1 \cdots X_{99}$. The IbE among these 100 agents would be *positive*, however, IbDp would be *zero* because the work that $X_1 \cdots X_{100}$ performs can be reduced to two disjoint parts, $X_1 \cdots X_{99}$ and X_{100}, working separately.

Example XorUnique (Figure 4.3) is analogous to the situation above. The whole specifies two bits of uncertainty, $\mathrm{I}(X_1X_2X_3 : Y) = \mathrm{H}(Y) = 2$ bits. The doublet X_1X_2 solely specifies the "digit-bit" of Y (0/1), $\mathrm{I}(X_1X_2 : Y) = 1$ bit, and the singleton X_3 solely specifies the "letter-bit" of Y (a/A), $\mathrm{I}(X_3 : Y) = 1$ bit. We apply each notion of irreducibility to XorUnique:

IbE How much of $X_1X_2X_3$'s information about Y can be reduced to the information conveyed by the singleton elements working separately? Working alone, X_3 still specifies the letter-bit of Y, but X_1 nor X_2 can unilaterally specify the digit-bit of Y, $\mathrm{I}(X_1 : Y) = 0$ and $\mathrm{I}(X_2 : Y) = 0$ bits. As only the letter-bit is specified when the three singletons work separately, $\mathsf{IbE}(\mathbf{X} : Y) = \mathrm{I}(X_1X_2X_3 : Y) - 1 = 2 - 1 = 1$ bit.

IbDp How much of $X_1X_2X_3$'s information about Y can be reduced to the information conveyed by disjoint parts working separately? Per subfigures 4.2g–i, there are three bipartitions of $X_1X_2X_3$, and one of them is $\{X_1X_2, X_3\}$. The doublet part X_1X_2 specifies the digit-bit of Y, and the singleton part X_3 specifies the letter-bit of Y. As there is a partition of $X_1X_2X_3$ that fully accounts for $X_1X_2X_3$'s specification of Y, $\mathsf{IbDp}(\mathbf{X} : Y) = 2 - 2 = 0$ bits.

Ib2p/IbAp How much of $X_1X_2X_3$'s information about Y can be reduced to the information conveyed by two parts working separately? From above we see that IbDp is zero bits. Per eq. (4.7), Ib2p and IbAp are stricter notions of irreducibility than IbDp, therefore Ib2p and IbAp must also be zero bits.

4.4.2 DoubleXor: Irreducible to a partition, yet reducible to a pair

In example DoubleXor (Figure 4.4), the whole specifies two bits, $\mathrm{I}(X_1X_2X_3 : Y) = \mathrm{H}(Y) = 2$ bits. The doublet X_1X_2 solely specifies the "left-bit", and the doublet X_2X_3 solely specifies the "right-bit". Applying each notion of irreducibility to DoubleXor:

IbE How much of $X_1X_2X_3$'s information about Y can be reduced to the information conveyed by singleton elements? The three singleton elements specify nothing about Y, $\mathrm{I}(X_i{:}Y) = 0$ bits $\forall i$. This means the whole is utterly irreducible to its elements, making $\mathsf{Ib}\mathsf{E}\,(\mathbf{X} : Y) = \mathrm{I}(X_1X_2X_3{:}Y) - 0 = 2$ bits.

IbDp How much of $X_1X_2X_3$'s information about Y can be reduced to the information conveyed by disjoint parts? Per subfigures 4.2g–i, the three bipartitions of $X_1X_2X_3$ are: $\{X_1X_2, X_3\}$, $\{X_1X_3, X_2\}$, and $\{X_2X_3, X_1\}$. In the first bipartition, $\{X_1X_2, X_3\}$, the doublet X_1X_2 specifies the left-bit of Y and the singleton X_3 specifies nothing for a total of one bit. Similarly, in the second bipartition, $\{X_2X_3, X_1\}$, X_2X_3 specifies the right-bit of Y and the singleton X_1 specifies nothing for a total of one bit. Finally, in the bipartition $\{X_2X_3, X_1\}$ both X_2X_3 and X_1 specify nothing for a total of zero bits. Taking the maximum over the three bipartitions, $\max[1, 1, 0] = 1$, we discover disjoint parts specify at most one bit, leaving $\mathsf{Ib}\mathsf{Dp}(\mathbf{X} : Y) = \mathrm{I}(X_1X_2X_3{:}Y) - 1 = 2 - 1 = 1$ bit.

Ib2p How much of $X_1X_2X_3$'s information about Y can be reduced to the information conveyed by two parts? Per subfigures 4.2k–j, there are three pairs of Almosts, and one of them is $\{X_1X_2, X_1X_3\}$. The Almost X_1X_2 specifies the left-bit of Y, and the Almost X_1X_3 specifies the right-bit of Y. As there is a pair of parts that fully accounts for $X_1X_2X_3$'s specification of Y, $\mathsf{Ib2p}(\mathbf{X} : Y) = 0$ bits.

IbAp How much of $X_1X_2X_3$'s information about Y can be reduced to the information conveyed by all possible parts? From above we see that Ib2p is zero bits. Per eq. (4.7), IbAp is stricter than Ib2p, therefore IbAp is also zero bits.

4.4.3 TripleXor: Irreducible to a pair of components, yet still reducible

Example TripleXor (Figure 4.5) has trifold symmetry and the whole specifies three bits, $\mathrm{I}(X_1X_2X_3{:}Y) = \mathrm{H}(Y) = 3$ bits. Each bit is solely specified by one of three doublets: X_1X_2, X_1X_3, or X_2X_3. Applying each notion of irreducibility to TripleXor:

IbE Working individually, the three elements specify absolutely nothing about Y, $\mathrm{I}(X_1{:}Y) = \mathrm{I}(X_2{:}Y) = \mathrm{I}(X_3{:}Y) = 0$ bits. Thus, the whole is utterly irreducible to elements, making $\mathsf{Ib}\mathsf{E}\,(\mathbf{X} : Y) = \mathrm{I}(X_1X_2X_3{:}Y) - 0 = 3$ bits.

IbDp The three bipartitions of $X_1X_2X_3$ are: $\{X_1X_2, X_3\}$, $\{X_1X_3, X_2\}$, and $\{X_2X_3, X_1\}$. In the first bipartition, $\{X_1X_2, X_3\}$, the doublet X_1X_2 specifies one bit of Y and the singleton X_3 specifies nothing for a total of one bit. By TripleXor's trifold symmetry, we get the same value for bipartitions $\{X_1X_2, X_3\}$ and $\{X_2X_3, X_1\}$. Taking the maximum over the three bipartitions,

$\max[1,1,1] = 1$, we discover a partition specifies at most one bit, leaving $\mathsf{IbDp}(\mathbf{X}:Y) = \mathrm{I}(X_1X_2X_3:Y) - 1 = 2$ bits.

Ib2p There are three pairs of Almosts: $\{X_1X_2, X_2X_3\}$, $\{X_1X_2, X_1X_3\}$, and $\{X_1X_3, X_2X_3\}$. Each pair of Almosts specifies exactly two bits. Taking the maximum over the pairs, $\max[2,2,2] = 2$, we discover a pair of parts specifies at most two bits, leaving $\mathsf{Ib2p}(\mathbf{X}:Y) = \mathrm{I}(X_1X_2X_3:Y) - 2 = 3 - 2 = 1$ bit.

IbAp The n Almosts of $X_1X_2X_3$ are $\{X_1, X_2, X_1X_3, X_2X_3\}$. Each Almost specifies one bit of Y, for a total of three bits, making $\mathsf{IbAp}(\mathbf{X}:Y) = \mathrm{I}(X_1X_2X_3:Y) - 3 = 0$ bits.

4.4.4 Parity: Complete irreducibility

In example PARITY (Figure 4.6), the whole specifies one bit of uncertainty, $\mathrm{I}(X_1X_2X_3:Y) = \mathrm{H}(Y) = 1$ bit. No singleton or doublet specifies anything about Y, $\mathrm{I}(X_i:Y) = \mathrm{I}\big(X_iX_j:Y\big) = 0$ bits $\forall i,j$. Applying each notion of irreducibility to PARITY:

IbE The whole specifies one bit, yet the elements $\{X_1, X_2, X_3\}$ specify nothing about Y. Thus the whole is utterly irreducible to elements, making $\mathsf{IbE}(\mathbf{X}:Y) = \mathrm{I}(X_1X_2X_3:Y) - 0 = 1$ bit.

IbDp The three bipartitions of $\mathbf{X}$ are: $\{X_1X_2, X_3\}$, $\{X_1X_3, X_2\}$, and $\{X_2X_3, X_1\}$. By the above each doublet and singleton specifies nothing about Y, and thus each partition specifies nothing about Y. Taking the maximum over the bipartitions yields $\max[0,0,0] = 0$, making $\mathsf{IbDp}(\mathbf{X}:Y) = 1 - 0 = 1$ bit.

Ib2p The pairs of $\mathbf{X}$'s Almosts are: $\{X_1X_2, X_1X_3\}$, $\{X_1X_2, X_2X_3\}$, and $\{X_1X_3, X_2X_3\}$. As before, each doublet specifies nothing about Y, and a pair of nothings is still nothing. Taking the maximum yields $\max[0,0,0] = 0$, making $\mathsf{Ib2p}(\mathbf{X}:Y) = 1 - 0 = 1$ bit.

IbAp The three Almosts of $\mathbf{X}$ are: $\{X_1X_2, X_1X_3, X_2X_3\}$. Each Almost specifies nothing, and a triplet of nothings is still nothing, making $\mathsf{IbAp}(\mathbf{X}:Y) = 1 - 0 = 1$ bit.

Table 4.1 summarizes the results of our four irreducibility measures applied to our examples.

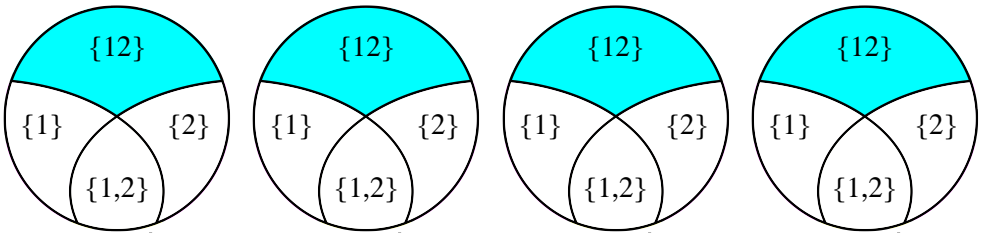
4.5 Conclusion

Within the Partial Information Decomposition framework[36], synergy is the simplest case of the broader notion of irreducibility. PI-diagrams, a generalization of Venn diagrams, are immensely helpful in improving one's intuition for synergy and irreducibility.

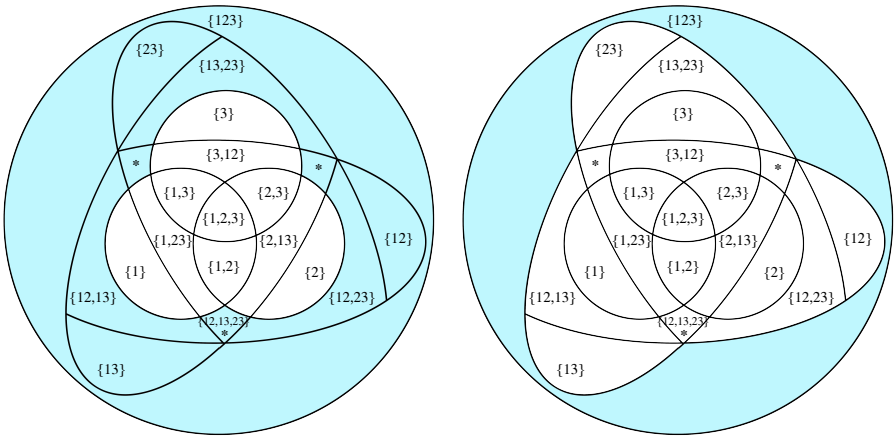
Example	$\mathrm{I}(X_{1\ldots n}:Y)$	IbE	IbDp	Ib2p	IbAp
RDN (Fig. 1.3)	1	0	0	0	0
UNQ (Fig. 1.4)	2	0	0	0	0
XOR (Fig. 1.5)	1	1	1	1	1
XORUNIQUE (Fig. 4.3)	2	1	0	0	0
DOUBLEXOR (Fig. 4.4)	2	2	1	0	0
TRIPLEXOR (Fig. 4.5)	3	3	2	1	0
PARITY (Fig. 4.6)	1	1	1	1	1

Table 4.1: Irreducibility values for our exemplary binary circuits.

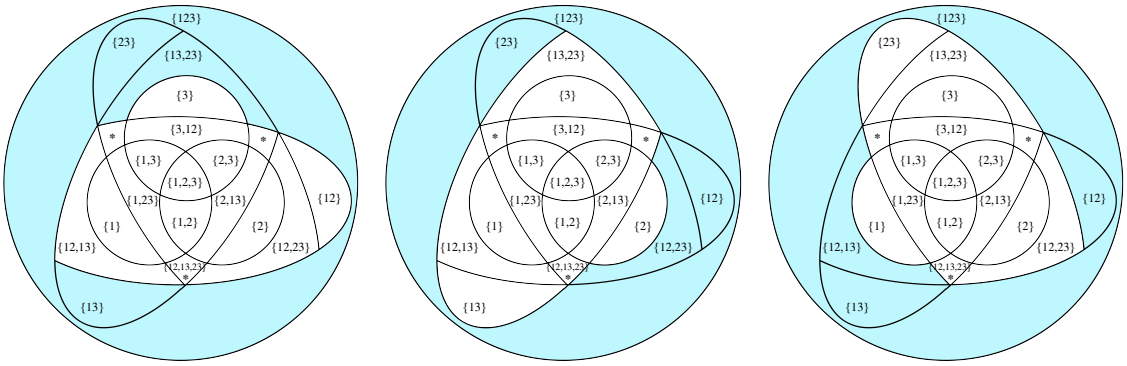
We define the irreducibility of the mutual information a set of n random variables $\mathbf{X} = \{X_1, \ldots, X_n\}$ convey about a target Y as the information the whole conveys about Y, $\mathrm{I}(X_{1\ldots n}:Y)$, minus the maximum union-information conveyed by the "parts" of $\mathbf{X}$. The four common notions of $\mathbf{X}$'s parts are: (1) the set of the n atomic elements; (2) all partitions of disjoint parts; (3) all pairs of parts; and (4) the set of all $2^n - 2$ possible parts. All four definitions of parts are equivalent when the whole consists of two atomic elements ($n = 2$), but they diverge for $n > 2$.



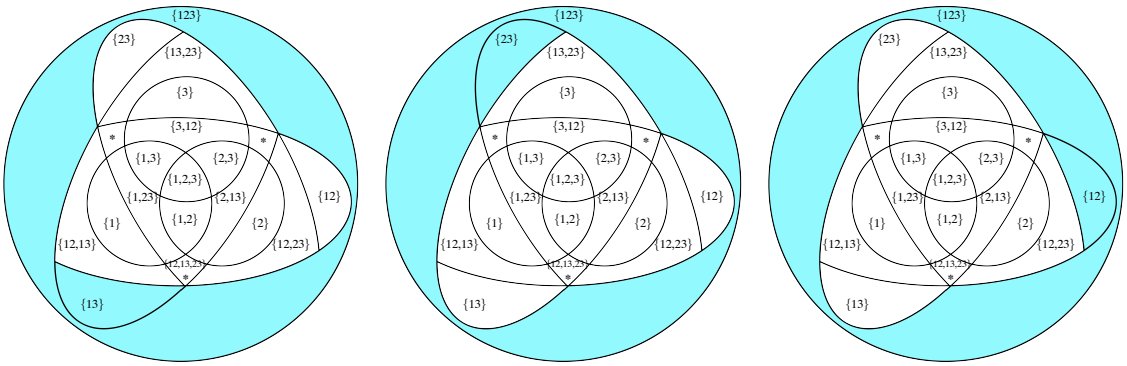
(a) $\mathsf{IbE}(X_1, X_2 : Y)$ (b) $\mathsf{IbDp}(X_1, X_2 : Y)$ (c) $\mathsf{Ib2p}(X_1, X_2 : Y)$ (d) $\mathsf{IbAp}(X_1, X_2 : Y)$



(e) $\mathsf{IbE}(X_1, X_2, X_3 : Y)$ (f) $\mathsf{IbAp}(X_1, X_2, X_3 : Y)$

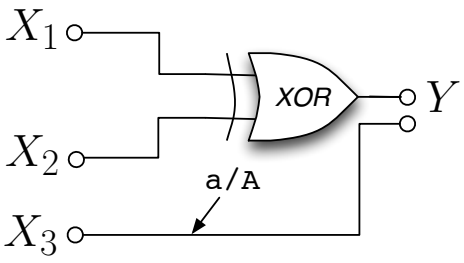


(g) $\mathbf{P} = \{X_1X_2, X_3\}$ (h) $\mathbf{P} = \{X_1X_3, X_2\}$ (i) $\mathbf{P} = \{X_2X_3, X_1\}$



(j) $\mathbf{P} = \{X_1X_2, X_2X_3\}$ (k) $\mathbf{P} = \{X_1X_2, X_1X_3\}$ (l) $\mathbf{P} = \{X_1X_3, X_2X_3\}$

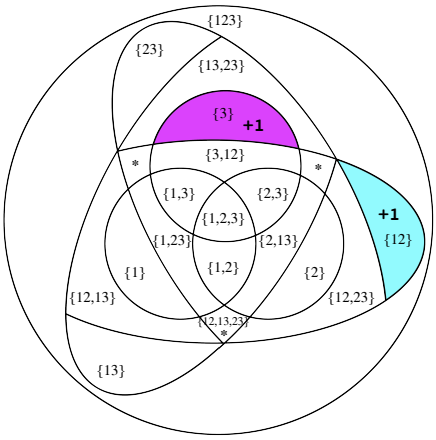
Figure 4.2: PI-diagrams depicting our four irreducibility measures when $n = 2$ and $n = 3$ in subfigures (a)–(d) and (e)–(l) respectively. For $n = 3$: IbE is (e), IbAp is (f), IbDp is the *minimum* value over subfigures (g)–(i), and $\mathsf{Ib2p}$ is the *minimum* value over subfigures (j)–(l).



(a) circuit diagram

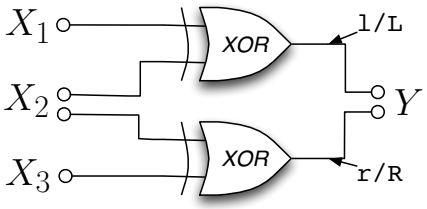
$X_1\,X_2\,X_3$	Y	
0 0 a	0a	1/8
0 1 a	1a	1/8
1 0 a	1a	1/8
1 1 a	0a	1/8
0 0 A	0A	1/8
0 1 A	1A	1/8
1 0 A	1A	1/8
1 1 A	0A	1/8

(b) $\Pr(x_1, x_2, x_3, y)$



(c) PI-diagram

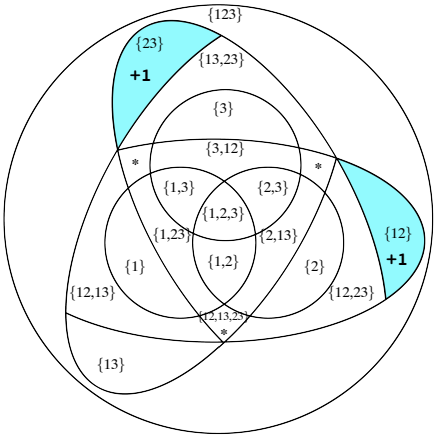
Figure 4.3: Example XorUnique. Target Y has two bits of uncertainty. The doublet X_1X_2 specifies the "digit bit", and the singleton X_3 specifies the "letter bit" for a total of $\mathrm{I}(X_1X_2{:}Y) + \mathrm{I}(X_3{:}Y) = \mathrm{H}(Y) = 2$ bits. $X_1X_2X_3$'s specification of Y is irreducible to singletons yet fully reduces to the disjoint parts $\{X_1X_2, X_3\}$.



(a) circuit diagram

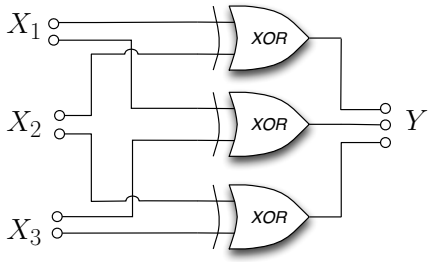
See Appendix 4.A for the joint distribution.

(b) $\Pr(x_1, x_2, x_3, y)$



(c) PI-diagram

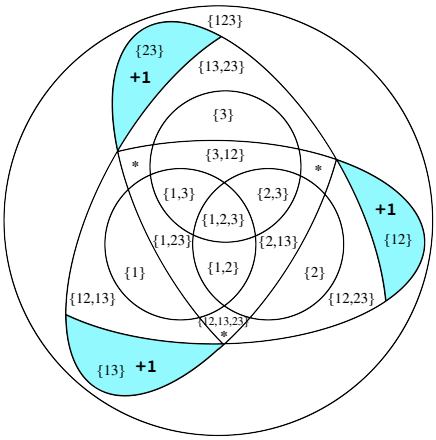
Figure 4.4: Example DoubleXor. Target Y has two bits of uncertainty. The doublet X_1X_2 specifies the "left bit" (l/L) and doublet X_2X_3 specifies the "right bit" (r/R) for a total of $\mathrm{I}(X_1X_2{:}Y) + \mathrm{I}(X_2X_3{:}Y) = \mathrm{H}(Y) = 2$ bits. $X_1X_2X_3$'s specification of Y is irreducible to disjoint parts yet fully reduces to the pair of parts $\{X_1X_2, X_2X_3\}$.



(a) circuit diagram

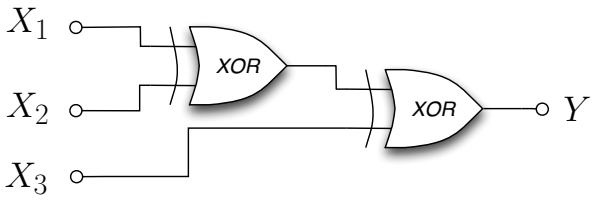
See Appendix 4.A for the joint distribution.

(b) $\Pr(x_1, x_2, x_3, y)$



(c) PI-diagram

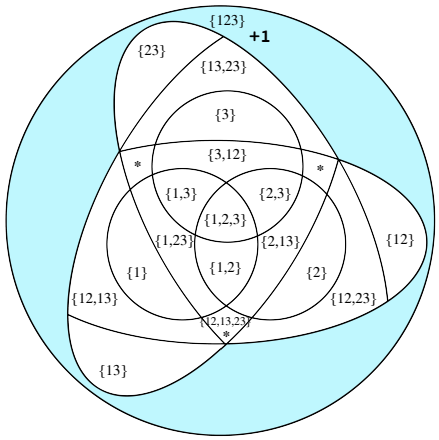
Figure 4.5: Example TripleXor. Target Y has three bits of uncertainty. Each doublet part of $X_1X_2X_3$ specifies a distinct bit of Y, for a total of $\mathrm{I}(X_1X_2:Y)+\mathrm{I}(X_1X_3:Y)+\mathrm{I}(X_2X_3:Y) = \mathrm{H}(Y) = 3$ bits. The whole's specification of Y is irreducible to any pair of Almosts yet fully reduces to all Almosts.



(a) circuit diagram

X_1	X_2	X_3	Y	
0	0	0	0	$1/8$
0	0	1	1	$1/8$
0	1	0	1	$1/8$
0	1	1	0	$1/8$
1	0	0	1	$1/8$
1	0	1	0	$1/8$
1	1	0	0	$1/8$
1	1	1	1	$1/8$

(b) $\Pr(x_1, x_2, x_3, y)$



(c) PI-diagram

Figure 4.6: Example Parity. Target Y has one bit of uncertainty, and only the whole specifies Y, $\mathrm{I}(X_1X_2X_3:Y) = \mathrm{H}(Y) = 1$ bit. $X_1X_2X_3$'s specification of Y is utterly irreducible to any collection of $X_1X_2X_3$'s parts, and $\mathsf{IbAp}(\{X_1, X_2, X_3\} : Y) = 1$ bit.

Appendix

4.A Joint distributions for DoubleXor and TripleXor

X_1	X_2	X_3	Y	
0	00	0	lr	$1/16$
0	01	0	lR	$1/16$
0	10	0	Lr	$1/16$
0	11	0	LR	$1/16$
0	00	1	lR	$1/16$
0	01	1	lr	$1/16$
0	10	1	LR	$1/16$
0	11	1	Lr	$1/16$
1	00	0	Lr	$1/16$
1	01	0	LR	$1/16$
1	10	0	lr	$1/16$
1	11	0	lR	$1/16$
1	00	1	LR	$1/16$
1	01	1	Lr	$1/16$
1	10	1	lR	$1/16$
1	11	1	lr	$1/16$

Figure 4.7: Joint distribution $\Pr(x_1, x_2, x_3, y)$ for example DoubleXor.

4.B Proofs

Lemma 4.B.1. *We prove that Information beyond the Bipartition,* $\mathit{Ib2p}(\mathbf{X} : Y)$*, equals Information beyond the Disjoint Parts,* $\mathit{IbDp}(\mathbf{X} : Y)$ *by showing,*

$$\mathit{IbDp}(\mathbf{X} : Y) \leq \mathit{Ib2p}(\mathbf{X} : Y) \leq \mathit{IbDp}(\mathbf{X} : Y) \ .$$

X_1	X_2	X_3	Y	
00	00	00	000	$1/64$
00	00	01	001	$1/64$
00	00	10	010	$1/64$
00	00	11	011	$1/64$
00	01	00	001	$1/64$
00	01	01	000	$1/64$
00	01	10	011	$1/64$
00	01	11	010	$1/64$
00	10	00	100	$1/64$
00	10	01	101	$1/64$
00	10	10	110	$1/64$
00	10	11	111	$1/64$
00	11	00	101	$1/64$
00	11	01	100	$1/64$
00	11	10	111	$1/64$
00	11	11	110	$1/64$
01	00	00	000	$1/64$
01	00	01	001	$1/64$
01	00	10	010	$1/64$
01	00	11	011	$1/64$
01	01	00	001	$1/64$
01	01	01	000	$1/64$
01	01	10	011	$1/64$
01	01	11	010	$1/64$
01	10	00	100	$1/64$
01	10	01	101	$1/64$
01	10	10	110	$1/64$
01	10	11	111	$1/64$
01	11	00	101	$1/64$
01	11	01	100	$1/64$
01	11	10	111	$1/64$
01	11	11	110	$1/64$

X_1	X_2	X_3	Y	
10	00	00	110	$1/64$
10	00	01	111	$1/64$
10	00	10	100	$1/64$
10	00	11	101	$1/64$
10	01	00	111	$1/64$
10	01	01	110	$1/64$
10	01	10	101	$1/64$
10	01	11	100	$1/64$
10	10	00	010	$1/64$
10	10	01	011	$1/64$
10	10	10	000	$1/64$
10	10	11	001	$1/64$
10	11	00	011	$1/64$
10	11	01	010	$1/64$
10	11	10	001	$1/64$
10	11	11	000	$1/64$
11	00	00	110	$1/64$
11	00	01	111	$1/64$
11	00	10	100	$1/64$
11	00	11	101	$1/64$
11	01	00	011	$1/64$
11	01	01	010	$1/64$
11	01	10	001	$1/64$
11	01	11	000	$1/64$
11	10	00	010	$1/64$
11	10	01	011	$1/64$
11	10	10	000	$1/64$
11	10	11	001	$1/64$
11	11	00	011	$1/64$
11	11	01	010	$1/64$
11	11	10	001	$1/64$
11	11	11	000	$1/64$

Figure 4.8: Joint distribution $\Pr(x_1, x_2, x_3, y)$ for example TripleXor.

Proof. We first show that $\mathsf{IbDp}(\mathbf{X} : Y) \leq \mathsf{Ib2p}(\mathbf{X} : Y)$. By their definitions:

$$\mathsf{IbDp}(\mathbf{X} : Y) \equiv \mathrm{I}(Y : X_{1\ldots n}) - \max_{\mathbf{P}} \mathrm{I}_{\cup}(Y : \mathbf{P}) \tag{4.8}$$

$$\mathsf{IbB}(\mathbf{X} : Y) \equiv \mathrm{I}(Y : X_{1\ldots n}) - \max_{S \subset \mathbf{X}} \mathrm{I}_{\cup}\left(Y : \{S, \mathbf{X} \setminus S\}\right) \tag{4.9}$$

$$= \mathrm{I}(Y : X_{1\ldots n}) - \max_{\substack{\mathbf{P} \\ |\mathbf{P}|=2}} \mathrm{I}_{\cup}(Y : \mathbf{P}) \ , \tag{4.10}$$

where $\mathbf{P}$ enumerates over all disjoint parts of $\mathbf{X}$.

By removing the restriction that $|\mathbf{P}| = 2$ from the minimized union-information in IbB, we arrive

at IbDp. As removing a restriction can only decrease the minimum, therefore $\mathsf{IbDp}(\mathbf{X} : Y) \leq \mathsf{IbB}(\mathbf{X} : Y)$. □

We next show that $\mathsf{IbB}(\mathbf{X} : Y) \leq \mathsf{IbDp}(\mathbf{X} : Y)$. Meaning we must show that,

$$\mathrm{I}(X_{1\ldots n} : Y) - \max_{\substack{\mathbf{P} \\ |\mathbf{P}|=2}} \mathrm{I}_{\cup}(\mathbf{P} : Y) \leq \mathrm{I}(X_{1\ldots n} : Y) - \max_{\mathbf{P}} \mathrm{I}_{\cup}(\mathbf{P} : Y) \ . \tag{4.11}$$

Proof. By subtracting $\mathrm{I}(X_{1\ldots n} : Y)$ from each side and multiplying each side by -1 we have,

$$\max_{\substack{\mathbf{P} \\ |\mathbf{P}|=2}} \mathrm{I}_{\cup}(\mathbf{P} : Y) \geq \max_{\mathbf{P}} \mathrm{I}_{\cup}(\mathbf{P} : Y) \ . \tag{4.12}$$

Without loss of generality, we take any individual subset/part S in X. Then we have a bipartition $\mathbf{B}$ of parts $\{S, \mathbf{X} \backslash S\}$. We then further partition the part $\mathbf{X} \backslash S$ into k disjoint subcomponents denoted $\{T_1, \ldots, T_k\}$ where $2 \leq k \leq n - |S|$, creating an arbitrary partition $\mathbf{P} = \{S, T_1, \ldots, T_k\}$. We now need to show that,

$$\mathrm{I}_{\cup}\Big(\big\{S, \mathbf{X} \setminus S\big\} : Y\Big) \geq \mathrm{I}_{\cup}\big(\{S, T_1, \ldots, T_k\} : Y\big) \ . \tag{4.13}$$

By the monotonicity axiom **(M)**, we can append each subcomponent $T_1, \ldots, T_k$ to $\mathbf{B}$ without changing the union-information, because every subcomponent T_i is a subset of the element $\mathbf{X} \setminus S$. Then, using the symmetry axiom ($\mathbf{S_0}$), we re-order the parts so that $S, T_1, \ldots, T_k$ come first. This yields,

$$\mathrm{I}_{\cup}\Big(\big\{S, T_1, \ldots, T_k, \mathbf{X} \setminus S\big\} : Y\Big) \geq \mathrm{I}_{\cup}\big(\{S, T_1, \ldots, T_k\} : Y\big) \ . \tag{4.14}$$

Applying the monotonicity axiom **(M)** again, we know that adding the entry $\mathbf{X} \setminus S$ can only increase the union information. Therefore, we prove eq. (5.13), which proves eq. (4.11). □

Finally, by the squeeze theorem we complete the proof of eq. (5.11), that $\mathsf{IbB}(\mathbf{X} : Y) = \mathsf{IbDp}(\mathbf{X} : Y)$.

Lemma 4.B.2. *Proof that pairs of Almosts cover* $\mathsf{Ib2p}$. *We prove that the maximum union-information over all possible pairs of parts* $\{P_1, P_2\}$, *equates to the maximum union-information over all pairs of Almosts* $\{A_i, A_j\}$ $i \neq j$. *Mathematically,*

$$\max_{\substack{P_1, P_2 \\ P_1, P_2 \subset \mathbf{X}}} \mathrm{I}_{\cup}\big(\{P_1, P_2\} : Y\big) = \max_{\substack{i,j \in \{1,\ldots,n\} \\ i \neq j}} \mathrm{I}_{\cup}\Big(\big\{A_i, A_j\big\} : Y\Big) \ . \tag{4.15}$$

Proof. By the right-monotonicity lemma **(RM)**, the union-information can only increase when increasing the size of the parts P_1 and P_2. We can therefore ignore all parts P_1, P_2 of size less than $n-1$,

$$\max_{\substack{P_1,P_2\\P_1,P_2\subset\mathbf{X}}} \mathrm{I}_\cup\big(\{P_1,P_2\}:Y\big) = \max_{\substack{P_1,P_2\\P_1,P_2\in\mathcal{P}(\mathbf{X})\\|P_1|=|P_2|=n-1}} \mathrm{I}_\cup\big(\{P_1,P_2\}:Y\big) \tag{4.16}$$

$$= \max_{i,j\in\{1,\ldots,n\}} \mathrm{I}_\cup\big(\{A_i,A_j\}:Y\big)\ . \tag{4.17}$$

Then by the idempotency axiom **(I)** and the monotonicity axiom **(M)**, having $i \neq j$ can only increase the union information. Therefore,

$$\max_{i,j\in\{1,\ldots,n\}} \mathrm{I}_\cup\big(\{A_i,A_j\}:Y\big) = \max_{\substack{i,j\in\{1,\ldots,n\}\\i\neq j}} \mathrm{I}_\cup\Big(\big\{A_i,A_j\big\}:Y\Big)\ . \tag{4.18}$$

With eq. (4.18) in hand, we easily show that the Information beyond all pairs of Subsets, $\mathsf{Ib2p}$, equates to the information beyond all pairs of Almosts:

$$\mathsf{Ib2p}\,(\mathbf{X}:Y) \equiv \mathrm{I}(X_{1\ldots n}:Y) - \max_{\substack{P_1,P_2\\P_1,P_2\in\mathcal{P}(\mathbf{X})}} \mathrm{I}_\cup\big(\{P_1,P_2\}:Y\big) \tag{4.19}$$

$$= \mathrm{I}(X_{1\ldots n}:Y) - \max_{\substack{i,j\in\{1,\ldots,n\}\\i\neq j}} \mathrm{I}_\cup\Big(\big\{A_i,A_j\big\}:Y\Big)\ . \tag{4.20}$$

□

Lemma 4.B.3. *Proof that Almosts cover IbAp. We wish to show that the union-information over all distinct parts of n elements, $\mathcal{P}(\mathbf{X})$, is equivalent to the union information over the n Almosts. Mathematically,*

$$\mathrm{I}_\cup\big(\mathcal{P}(\mathbf{X}):Y\big) = \mathrm{I}_\cup\big(\{A_1,\ldots,A_n\}:Y\big)\ . \tag{4.21}$$

Proof. Every element in the set of parts $\mathcal{P}(\mathbf{X})$ that isn't an Almost is a subset of an Almost. Therefore, by the monotonicity axiom **(M)** we can remove this entry. Repeating this process, we remove all entries except the n Almosts. Therefore, $\mathrm{I}_\cup\big(\mathcal{P}(\mathbf{X}):Y\big) = \mathrm{I}_\cup\big(\{A_1,\ldots,A_n\}:Y\big)$. □

Part III

Applications

Chapter 5

Improving the ϕ Measure

5.1 Introduction

The measure of integrated information, ϕ, is an attempt to a quantify the magnitude of conscious experience. It has a long history [31, 3, 33], and at least three different measures have been called ϕ. Here we consider some adjustments to the ϕ measure from [3] to correct perceived deficiencies.[1]

The ϕ measure aims to quantify a system's "functional irreducibility to disjoint parts." As discussed in Chapter 4, we can use Partial Information Decomposition (PID) to derive a principled measure of irreducibility to disjoint parts. This revised measure, ψ, has numerous desirable properties over ϕ.

5.2 Preliminaries

5.2.1 Notation

We use the following notation throughout this chapter:

n: the number of indivisible elements in network X. $n \geq 2$.

$\mathbf{P}$: a partition of the n indivisible nodes clustered into m parts. Each part has at least one node and each partition has at least two parts, so $2 \leq m \leq n$.

$X_i^{\mathbf{P}}$: a random variable representing a part i at time =0. $1 \leq i \leq m$.

$Y_i^{\mathbf{P}}$: a random variable representing part i after t updates. $1 \leq i \leq m$.

X: a random variable representing the entire network at time=0. $X \equiv X_1^{\mathbf{P}} \cdots X_m^{\mathbf{P}}$.

[1] We chose the 2008 version [3] because it is the most recent purely information-theoretic ϕ. The most recent version from [33] uses the Hamming Distance among states and thus changes depending on the chosen labels. We are aware of no other info-theoretic measure that changes under relabeling. Secondly, the measure in [33] is in units bits-squared, which has no known information-theoretic interpretation.

Y: a random variable representing the entire network after t applications of the neural network's update rule. $Y \equiv Y_1^{\mathbf{P}} \cdots Y_m^{\mathbf{P}}$.

y: a single state of the random variable Y.

X: The set of n indivisible elements at time=0.

For readers accustomed to the notation in [3], the translation is $X \equiv X_0$, $Y \equiv X_1$, $X_i^{\mathbf{P}} \equiv M_0^i$, and $Y_i^{\mathbf{P}} \equiv M_1^i$.

For pedagogical purposes we confine this paper to deterministic networks. Therefore all remaining entropy at time t conveys information about the past, i.e., $\mathrm{I}(X:Y) = \mathrm{H}(Y)$ and $\mathrm{I}\big(X:Y_i^{\mathbf{P}}\big) = \mathrm{H}\big(Y_i^{\mathbf{P}}\big)$ where $\mathrm{I}(\bullet:\bullet)$ is the mutual information and $\mathrm{H}(\bullet)$ is the Shannon entropy[9]. Our model generalizes to probabilistic units with any finite number of discrete—but not continuous—states[5]. All logarithms are $\log_2$. All calculations are in *bits*.

5.2.2 Model assumptions

(A) The ϕ measure is a *state-dependent* measure, meaning that every output state $y \in Y$ has its own ϕ value. To simplify cross-system comparisons, some researchers[5] prefer to consider only the averaged ϕ, denoted $\langle\phi\rangle$. Here we adhere to the original theoretical state-dependent formulation. But when comparing large numbers of networks we use $\langle\phi\rangle$ for convenience.

(B) The ϕ measure aims to quantify "information intrinsic to the system". This is often thought to be synonymous with causation, but it's not entirely clear. But for this reason, in [3] all random variables at time=0, i.e. X and $X_1^{\mathbf{P}}, \ldots, X_m^{\mathbf{P}}$ are made to follow an *independent discrete uniform distribution*. There are actually several plausible choices for the distribution on X , but for easier comparison to [3], here we also take X to be an independent discrete uniform distribution. This means that $\mathrm{H}(X) = \log_2 |X|$, $\mathrm{H}\big(X_i^{\mathbf{P}}\big) = \log_2 \big|X_i^{\mathbf{P}}\big|$ where $|\bullet|$ is the number of states in the random variable, and $\mathrm{I}\Big(X_i^{\mathbf{P}}:X_j^{\mathbf{P}}\Big) = 0\ \forall i \neq j$.

(C) We set $t = 1$, meaning we compute these informational measures for a system undergoing a single update from time=0 to time=1. This has no impact on generality (see Appendix 5.E). To analyze real biological networks one would sweep t over all reasonable timescales, choosing the t that maximizes the complexity metric.

5.3 How ϕ works

The ϕ measure has four steps and proceeds as follows:

1. For a given state $y \in Y$, [3] first defines the state's *effective information*, quantifying the total magnitude of information the state y conveys about X, the r.v. representing a maximally

ignorant past. This turns out to be identical to the "specific-surprise" measure, $\mathrm{I}(X\!:\!y)$, from [10],

$$\texttt{ei}(X \to y) = \mathrm{I}(X\!:\!y) = \mathrm{D}_{\mathrm{KL}}\left[\Pr\big(X|y\big)\middle\|\Pr(X)\right] \ . \tag{5.1}$$

Given X follows a discrete uniform distribution (assumption **(B)**), $\texttt{ei}(X \to y)$ simplifies to,

$$\begin{aligned}\texttt{ei}(X \to y) &= \mathrm{H}(X) - \mathrm{H}\big(X|y\big) \\ &= \mathrm{H}(X) - \sum_{x \in X} \Pr\big(x|y\big) \log \frac{1}{\Pr\big(x|y\big)} \ .\end{aligned} \tag{5.2}$$

In the nomenclature of [20], $\texttt{ei}(X \to y)$ can be understood as the "total causal power" the system exerts when transitioning into state y.

2. The ϕ measure aims to quantify a system's irreducibility to disjoint parts, and the second step is to quantify how much of the total causal power isn't accounted for by the disjoint parts (partition) $\mathbf{P}$. To do this, they define the *effective information beyond partition* $\boldsymbol{P}$,

$$\texttt{ei}\left(X \to y/\mathbf{P}\right) \equiv \mathrm{D}_{\mathrm{KL}}\left[\Pr(X|y)\middle\|\prod_{i=1}^{m}\Pr\Big(X_i^{\mathbf{P}}\Big|y_i^{\mathbf{P}}\Big)\right] \ . \tag{5.3}$$

The intuition behind $\texttt{ei}(X \to y/\mathbf{P})$ is that it quantifies the amount of causal power in $\texttt{ei}(X \to y)$ that is irreducible to the parts $\mathbf{P}$ operating independently.[2]

3. After defining the causal power beyond an arbitrary partition $\mathbf{P}$, the third step is to find the partition that accounts for as much causal power as possible. This partition is called the *Minimum Information Partition*, or MIP. They define the MIP for a given state y as,[3]

$$\mathrm{MIP}(y) \equiv \operatorname*{argmin}_{\mathbf{P}} \frac{\texttt{ei}(X \to y/\mathbf{P})}{(m-1)\cdot \min_i \mathrm{H}\big(X_i^{\mathbf{P}}\big)} \ . \tag{5.4}$$

Finding the MIP of a system by brute force is incredibly computationally expensive, as it requires enumerating all partitions of n nodes scales $O(n!)$ and even for supercomputers becomes intractable for $n > 32$ nodes.

4. Fourth and finally, the system's causal irreducibility when transitioning into state $y \in Y$, $\phi(y)$, is the effective information beyond y's MIP,

$$\phi(y) \equiv \texttt{ei}\left(X \to y\big/\mathbf{P} = \mathrm{MIP}(y)\right) \ .$$

[2] In [3] they deviated slightly from this formulation, using a process termed "perturbing the wires". However, subsequent work[32, 33] disavowed perturbing the wires and thus we don't use it here. For discussion see Appendix 5.C

[3] In [3] they additionally consider the *total partition* as a special case. However, subsequent work[32, 33] disavowed the total partition and thus we don't use it here.

5.3.1 Stateless ϕ is $\langle\phi\rangle$

In [3] ϕ is defined for every state $y \in Y$, and a single system can have a wide range of ϕ-values. In [5], they found this medley of state-dependent ϕ-values unwieldy, and they decided to get a single number per system by averaging the effective information over all states y. This gives rise to the four corresponding stateless measures:

$$\begin{aligned}
\big\langle \mathtt{ei}(Y) \big\rangle &\equiv \mathbb{E}_y \mathtt{ei}(X \to y) = \mathrm{I}(X\!:\!Y) \\
\big\langle \mathtt{ei}(Y/\mathbf{P}) \big\rangle &\equiv \mathbb{E}_y \mathtt{ei}\,(X \to y/\mathbf{P}) = \mathrm{I}(X\!:\!Y) - \sum_{i=1}^{m} \mathrm{I}\Big(X_i^{\mathbf{P}}\!:\!Y_i^{\mathbf{P}}\Big) \\
\langle \mathrm{MIP} \rangle &\equiv \operatorname*{argmin}_{\mathbf{P}} \frac{\big\langle \mathtt{ei}(Y/\mathbf{P}) \big\rangle}{(m-1)\cdot \min_i \mathrm{H}\big(X_i^{\mathbf{P}}\big)} \\
\langle \phi \rangle &\equiv \Big\langle \mathtt{ei}\left(Y/\mathbf{P} = \langle \mathrm{MIP} \rangle\right) \Big\rangle \ .
\end{aligned} \tag{5.5}$$

Although the distinction has yet to affect qualitative results, researchers should note that $\langle\phi\rangle \neq \mathbb{E}_Y \phi(y)$. This is because whereas each y state can have a different MIP, for $\langle\phi\rangle$ there's only one MIP for all states.

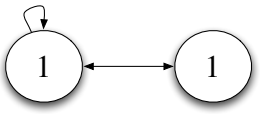
5.4 Room for improvement in ϕ

$\phi(y)$ can exceed $\mathrm{H}(X)$. Figure 5.1 shows examples OR-GET and OR-XOR. On average, each looks fine—they each have $\mathrm{H}(X) = 2$, $\mathrm{I}(X\!:\!Y) = 1.5$, and $\langle\phi\rangle = 1.189$ bits—nothing peculiar. This changes when examining the individual states $y \in Y$.

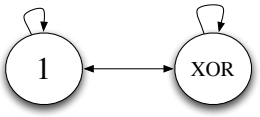
For OR-GET, the $\phi(y = \mathtt{10}) \approx 2.58$ bits. Therefore $\phi(y)$ *exceeds* the entropy of the entire system, $\mathrm{H}(XY) = \mathrm{H}(X) = 2$ bits. This means that for $y = \mathtt{10}$, the "irreducible causal power" exceeds not just the total causal power, $\mathtt{ei}(X \to y)$, but $\mathtt{ei}$'s upperbound of $\mathrm{H}(X)$! This is concerning.

For OR-XOR, $\phi(y = \mathtt{11}) \approx 1.08$ bits. This does not exceed $\mathrm{H}(X)$, but it does exceed $\mathrm{I}(X\!:\!y = \mathtt{11}) = 1$ bit. Per eq. (5.5), in expectation $\big\langle \mathtt{ei}(Y/\mathbf{P}) \big\rangle \leq \mathrm{I}(X\!:\!Y)$ for any partition $\mathbf{P}$. An information-theoretic interpretation of the state-dependent case would be more natural if likewise $\mathtt{ei}(X \to y/\mathbf{P}) \leq \mathrm{I}(X\!:\!Y = y)$ for any partition $\mathbf{P}$. Note this issue is not due simply to normalizing in eq. (5.4). For OR-GET and OR-XOR there's only one possible partition, and thus the normalization has no effect. The oddity arises from the equation for the effective information beyond a partition, eq. (5.3).

ϕ sometimes decreases with duplicate computation. In Figure 5.2 we take a simple system, AND-ZERO, and duplicate the AND gate yielding AND-AND. We see the two systems remain exceedingly similar. Each has $\mathrm{H}(X) = 2$ and $\mathrm{I}(X\!:\!Y) = 0.811$ bits. Likewise, each has two Y states occurring with probability $3/4$ and $1/4$, giving $\mathtt{ei}(X \to y)$ equal to 0.42 and 2.00 bits, respectively. However, their ϕ values are quite different.



(a) OR-GET network



(b) OR-XOR network

X		OR-GET	OR-XOR
00	→	00	00
01	→	10	11
10	→	11	11
11	→	11	10

Transition table for (a), (b)

	OR-GET (a)				OR-XOR (b)			
	00	01	10	11	00	01	10	11
Pr(y)	1/4	-	1/4	1/2	1/4	-	1/4	1/2
ei(y)	2.00	-	2.00	1.00	2.00	-	2.00	1.00
$\phi(y)$	1.00	-	**2.58**	0.58	1.00	-	1.58	**1.08**

Figure 5.1: Example OR-GET shows that $\phi(y)$ can exceed not only ei($X \rightarrow y$), but H(X)! A dash means that particular y is unreachable for the network. The concerning ϕ values are **bolded**.

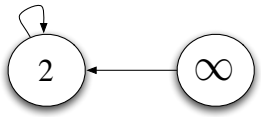
If we only knew that the ϕ's for AND-AND and AND-ZERO were different, we'd expect AND-AND's ϕ to be higher because an AND node "does more" than a ZERO node (simply shutting off). But instead we get the opposite—AND-AND's highest ϕ is *less* than AND-ZERO's lowest ϕ! An ideal measure of integrated information might be invariant or increase with duplicate computation, but it certainly wouldn't decrease!

ϕ does not increase with cooperation among diverse parts. The ϕ measure is often thought of as corresponding to the juxtaposition of "functional segregation" and "functional integration". In a similar vein, ϕ is also intuited as corresponding to "interdependence/cooperation among diverse parts". Figure 5.3 presents four examples showing that these similar intuitions are not captured by the existing ϕ measure.

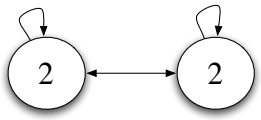
In the first example, SHIFT (Figure 5.3a), each bit one step clockwise—nothing more, nothing less. The nodes are homogeneous and each node is determined by its preceding node.

In the three remaining networks (Figures 5.3b–d), each node is a function of all nodes in its network (including itself). This is to maximize interdependence among the nodes, making the network highly "functionally integrated". Having established high cooperation, we increase the diversity/"functional segregation" from Figure 5.3b to 5.3d.

By the aforementioned intuitions, we'd expect SHIFT (Figure 5.3a) to have the lowest ϕ and 4321 (Figure 5.3d) to have the highest. But this is not the case. Instead, SHIFT, the network with the *least* cooperation (every node is a function of one other) and the *least* diverse mechanisms (all nodes have threshold 1) has a ϕ far exceeding the others; SHIFT's lowest ϕ value at 2.00 bits dwarfs the ϕ values in Figures 5.3b–d.



(a) AND-ZERO network



(b) AND-AND network

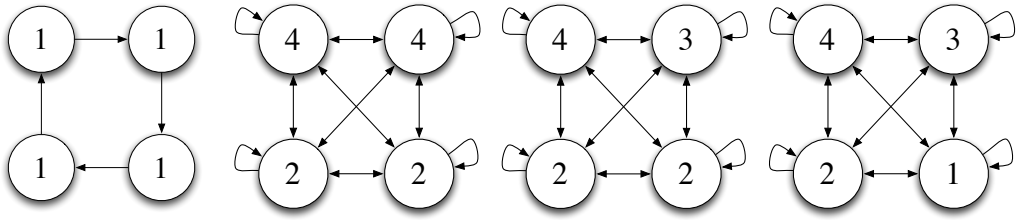
X		AND-ZERO	AND-AND
00	→	00	00
01	→	00	00
10	→	00	00
11	→	10	11

Transition table for (a), (b)

	AND-ZERO (a)				AND-AND (b)			
	00	01	10	11	00	01	10	11
$\Pr(y)$	3/4	-	1/4	-	3/4	-	-	1/4
$\texttt{ei}(y)$	0.42	-	2.00	-	0.42	-	-	2.00
$\phi(y)$	0.33	-	1.00	-	0.25	-	-	0.00

Figure 5.2: Examples AND-ZERO and AND-AND show that $\phi(y)$ sometimes *decreases* with duplicate computation. Here, the highest ϕ of AND-AND is *less* than the lowest ϕ of AND-ZERO. This carries into the average case with AND-ZERO's $\langle\phi\rangle = 0.5$ and AND-AND's $\langle\phi\rangle = 0.189$ bits. A dash means that particular y is unreachable for the network.

SHIFT having the highest ϕ is unexpected, but it's not outright absurd. In SHIFT each node is wholly determined by an external force (the preceding node); so in some sense SHIFT is "integrated". Whether it makes sense for SHIFT to have the highest integrated information ultimately comes down to precisely what is meant by the term "integration". But even accepting that SHIFT is in some sense integrated, network 4321 is integrated for a stronger sense of the term. Therefore, until there's some argument that the awareness of SHIFT *should* be higher than 4321, from a purely theoretical perspective it makes sense to prefer 4321 over SHIFT.



(a) SHIFT (b) 4422 (c) 4322 (d) 4321

Network	$\mathrm{I}(X:Y)$	$\min_y \phi(y)$	$\max_y \phi(y)$	$\langle\phi\rangle$
SHIFT	4.000	2.000	2.000	2.000
4422	1.198	0.000	0.673	0.424
4322	1.805	0.322	1.586	1.367
4321	2.031	0.322	1.682	1.651

Figure 5.3: State-dependent ϕ and $\langle\phi\rangle$ tell the same story—the ϕ value of SHIFT (a) trounces the ϕ of the other three networks. Neither ϕ measure is representative of cooperation among diverse parts.

5.5 A Novel Measure of Irreducibility to a Partition

Our proposed measure ψ quantifies the magnitude of information in $\mathrm{I}(X:y)$ (eq. (5.1)) that is irreducible to a partition of the system at time=0. We define our measure as,

$$\psi(\mathbf{X}:y) \equiv \mathrm{I}(X:y) - \max_{\mathbf{P}} \mathrm{I}_{\cup}\Big(X_1^{\mathbf{P}}, \ldots, X_m^{\mathbf{P}}:y\Big) \ , \tag{5.6}$$

where $\mathbf{P}$ enumerates over all partitions of set $\mathbf{X}$, and $\mathrm{I}_{\cup}$ is the information about state y conveyed by the "union" across the m parts at time=0. To compute the union information $\mathrm{I}_{\cup}$ we use the Partial Information Decomposition (PID) framework. In PID, $\mathrm{I}_{\cup}$ is the dual to $\mathrm{I}_{\cap}$; they are related by the inclusion–exclusion principle. Thus we can express $\mathrm{I}_{\cup}$ by,

$$\mathrm{I}_{\cup}\Big(X_1^{\mathbf{P}}, \ldots, X_m^{\mathbf{P}}:y\Big) = \sum_{\mathbf{S}\subseteq\left\{X_1^{\mathbf{P}},\ldots,X_m^{\mathbf{P}}\right\}} (-1)^{|\mathbf{S}|+1}\, \mathrm{I}_{\cap}\Big(S_1, \ldots, S_{|\mathbf{S}|}:y\Big) \ .$$

Conceptually, the intersection information $\mathrm{I}_{\cap}\Big(S_1, \ldots, S_{|\mathbf{S}|}:y\Big)$ is the magnitude of information about state y that is conveyed "redundantly" by each $S_i \in \mathbf{S}$. Although there currently remains some debate[7, 14] about what is the best measure of $\mathrm{I}_{\cap}$, there's consensus that the intersection information n arbitrary random variables $Z_1, \ldots, Z_n$ convey about state $Y = y$ must satisfy the following properties:

(**GP**) Global Positivity: $\mathrm{I}_{\cap}(Z_1, \ldots, Z_n:y) \geq 0$ with equality if $\Pr(y) = 0$ or $\Pr(y) = 1$.

(**$\mathbf{M_0}$**) Weak Monotonicity: $\mathrm{I}_{\cap}(Z_1, \ldots, Z_n, W:y) \leq \mathrm{I}_{\cap}(Z_1, \ldots, Z_n:y)$ with equality if there exists $Z_i \in \{Z_1, \ldots, Z_n\}$ such that $\mathrm{H}(Z_i|W) = 0$.

(**$\mathbf{S_0}$**) Weak Symmetry: $\mathrm{I}_{\cap}(Z_1, \ldots, Z_n:y)$ is invariant under reordering $Z_1, \ldots, Z_n$.

(**SR**) Self-Redundancy: $\mathrm{I}_{\cap}(Z_1:y) = \mathrm{I}(Z_1:y) = \mathrm{D}_{\mathrm{KL}}\Big[\Pr\big(Z_1|y\big)\Big\|\Pr(Z_1)\Big]$. The intersection information a single predictor Z_1 conveys about the target state $Y = y$ is equal to the "specific surprise"[10] between the predictor and the target state.

(**Eq**) Equivalence-Class Invariance: $\mathrm{I}_{\cap}(Z_1, \ldots, Z_n:y)$ is invariant under substituting Z_i (for any $i = 1, \ldots, n$) by an informationally equivalent random variable[4].[14] Similarly, $\mathrm{I}_{\cap}(Z_1, \ldots, Z_n:y)$ is invariant under substituting state y for state w if $\Pr\big(w|y\big) = \Pr\big(y|w\big) = 1$.

Instead of choosing a particular $\mathrm{I}_{\cap}$ that satisfies the above properties, we will simply use these properties directly to bound the range of possible ψ values. Leveraging (**$\mathbf{M_0}$**), (**$\mathbf{S_0}$**), and (**SR**),

[4]Meaning that $\mathrm{I}_{\cap}$ is invariant under substituting Z_i with W if $\mathrm{H}\big(Z_i|W\big) = \mathrm{H}\big(W|Z_i\big) = 0$.

eq. (5.6) simplifies to,[5]

$$\begin{aligned}\psi(\mathbf{X} : y) &= \mathrm{I}(X : y) - \max_{A \subset \mathbf{X}} \mathrm{I}_\cup(A, B : y) \\ &= \mathrm{I}(X : y) - \max_{A \subset \mathbf{X}} \left[\mathrm{I}(A : y) + \mathrm{I}(B : y) - \mathrm{I}_\cap(A, B : y)\right] \ ,\end{aligned} \tag{5.7}$$

where $A \neq \emptyset$ and $B \equiv \mathbf{X} \setminus A$.

From eq. (5.7), the only term left to be defined is $\mathrm{I}_\cap(A, B : y)$. Leveraging (**GP**), ($\mathbf{M_0}$), and (**SR**), we can bound this by $0 \leq \mathrm{I}_\cap(A, B : y) \leq \min\left[\mathrm{I}(A : y), \mathrm{I}(B : y)\right]$.

Finally, we bound ψ by plugging in the above bounds on $\mathrm{I}_\cap(A, B : y)$ into eq. (5.7). With some algebra and leveraging assumption **(B)**, this becomes,[6]

$$\begin{aligned}\psi_{\min}(\mathbf{X} : y) &= \min_{A \subset \mathbf{X}} \mathrm{D}_{\mathrm{KL}}\left[\Pr(X_{1\ldots n}|y) \middle\| \Pr\big(A|y\big) \Pr\big(B|y\big)\right] \\ \psi_{\max}(\mathbf{X} : y) &= \min_{i \in \{1,\ldots,n\}} \mathrm{D}_{\mathrm{KL}}\left[\Pr(X_{1\ldots n}|y) \middle\| \Pr(X_i) \Pr(X_{\sim i}|y)\right] \ ,\end{aligned} \tag{5.8}$$

where $X_{\sim i}$ is the random variable of all nodes in X excluding node i. Then, $\psi_{\min}(\mathbf{X} : y) \leq \psi(\mathbf{X} : y) \leq \psi_{\max}(\mathbf{X} : y)$.

5.5.1 Stateless ψ is $\langle\psi\rangle$

Matching how $\langle\phi\rangle$ is defined in Section 5.3.1, to compute $\langle\psi\rangle$ we weaken the properties in Section 5.5 so that they only apply to the average case, i.e., the properties (**GP**), ($\mathbf{M_0}$), ($\mathbf{S_0}$), (**SR**), and (**Eq**) don't have to apply for each $\mathrm{I}_\cap(Z_1, \ldots, Z_n : y)$, but merely for the average case $\mathrm{I}_\cap(Z_1, \ldots, Z_n : Y)$.

Via the same algebra[7], $\langle\psi\rangle$ simplifies to,

$$\begin{aligned}\langle\psi\rangle(X_1, \ldots, X_n : Y) &\equiv \mathrm{I}(X : Y) - \max_{\mathbf{P}} \mathrm{I}_\cup\left(X_1^{\mathbf{P}}, \ldots, X_m^{\mathbf{P}} : Y\right) \\ &= \mathrm{I}(X : Y) - \max_{A \subset \mathbf{X}} \mathrm{I}_\cup(A, B : Y) \\ &= \mathrm{I}(X : Y) - \max_{A \subset \mathbf{X}} \left[\mathrm{I}(A : Y) + \mathrm{I}(B : Y) - \mathrm{I}_\cap(A, B : Y)\right] \ ,\end{aligned} \tag{5.9}$$

where $A \neq \emptyset$ and $B \equiv \mathbf{X} \setminus A$. Using the weakened properties, we have $0 \leq \mathrm{I}_\cap(A, B : Y) \leq \min\left[\mathrm{I}(A : Y), \mathrm{I}(B : Y)\right]$. Plugging in these $\mathrm{I}_\cap$ bounds, we achieve the analogous bounds on $\langle\psi\rangle$,[8]

$$\begin{aligned}\langle\psi\rangle_{\min}(\mathbf{X} : Y) &= \min_{A \subset \mathbf{X}} \mathrm{I}(A : B|Y) \\ \langle\psi\rangle_{\max}(\mathbf{X} : Y) &= \min_{i \in \{1,\ldots,n\}} \mathrm{D}_{\mathrm{KL}}\left[\Pr(X, Y) \middle\| \Pr(X_{\sim i}, Y) \Pr(X_i)\right] \ ,\end{aligned} \tag{5.10}$$

[5] See Appendix 5.B.1 for a proof.
[6] See Appendix 5.B.2 for proofs.
[7] See Appendix 5.B.1 for a proof.
[8] See Appendix 5.B.3 for proofs.

where $X_{\sim i}$ is the random variable of all nodes in X excluding node i. Then,
$\langle\psi\rangle_{\min}(\mathbf{X}:Y) \leq \langle\psi\rangle(\mathbf{X}:Y) \leq \langle\psi\rangle_{\max}(\mathbf{X}:Y)$.

5.6 Contrasting ψ versus ϕ

Theoretical benefits. The overarching theoretical benefit of ψ is that it is entrenched within the rigorous Partial Information Decomposition framework[36]. PID builds a theoretically principled irreducibility measure from a redundancy measure $\mathrm{I}_\cap$. Here we only take the most accepted properties of $\mathrm{I}_\cap$ to bound ψ from above and below. As the complexity community converges on the additional properties $\mathrm{I}_\cap$ must satisfy[7, 14], the derived bounds on ψ will contract.

The first benefit of ψ's principled underpinning is that whereas $\phi(y)$ can exceed the entropy of the whole system, i.e., $\phi(y) \not\leq \mathrm{H}(X)$, $\psi(y)$ is bounded by specific-surprise, i.e., $\psi(y) \leq \mathrm{I}(X:y) = \mathrm{D}_{\mathrm{KL}}\left[\Pr(X|y)\middle\|\Pr(X)\right]$. This gives ψ the natural info-theoretic interpretation for the state-dependent case which ϕ lacks. A second benefit is that PID provides justification for ψ not needing a MIP normalization, and thus eliminates a longstanding concern about ϕ[2]. The third benefit is that PID is a flexible framework that enables quantifying irreducibility to overlapping parts should we decide to explore it[9].

One final perk is that ψ is substantially faster to compute. Whereas computing ϕ scales[10] $O(n!)$, computing ψ scales[11] $O(2^n)$—a substantial improvement that may improve even further as the complexity community converges on additional properties of $\mathrm{I}_\cap$.

Practical differences. The first row in Figure 5.4 shows two ways a network can be irreducible to atomic elements (the nodes) yet still reducible to disjoint parts. Compare AND-ZERO (Figure 5.4g) to AND-ZERO+KEEP (Figure 5.4a). Although AND-ZERO is irreducible, AND-ZERO+KEEP reduces to the bipartition separating the AND-ZERO component and the KEEP node. This reveals how fragile measures like ψ and ϕ are—add a single disconnected node and they plummet to zero. Example 2x AND-ZERO (Figure 5.4b) shows that a reducible system can be composed entirely of irreducible parts.

Example KEEP-KEEP (Figure 5.4c) highlights the only known relative drawback of ψ: its current upperbound[12] is painfully loose. The desired irreducibility for KEEP-KEEP is zero bits, and indeed, $\psi_{\min}$ is 0 bits, but $\psi_{\max}$ is a monstrous 1 bit! We rightly expect tighter bounds for such easy examples like KEEP-KEEP. Tighter bounds on $\mathrm{I}_\cap$ (and thus ψ) is an area of active research but as-is the bounds are loose.

Example GET-GET (Figure 5.4d) epitomizes the most striking difference between ψ and ϕ.

[9] Technically there are multiple irreducibilities to overlapping parts as, unlike disjoint parts, the maximum union information over two overlapping parts is not equal to the maximum union information over m overlapping parts.
[10] This comes from eq. (5.4) enumerating all partitions (Bell's number) of n elements.
[11] This comes from eq. (5.7) enumerating all $2^{n-1}-1$ bipartitions of n elements.
[12] The current upperbounds are $\psi_{\max}$ in eq. (5.8) and $\langle\psi\rangle_{\max}$ in eq. (5.10).

By property (**Eq**), the ψ values for KEEP-KEEP and GET-GET are provably equal (making the desired ψ for GET-GET also zero bits), yet their ϕ values couldn't be more different. Although ϕ agrees KEEP-KEEP is zero, ϕ firmly places GET-GET at the maximal (!) two bits of irreducibility. Whereas ψ views GET nodes akin to a system-wide KEEP, ϕ views GET nodes as highly integrative.

The primary benefit of making KEEPs and GETs equivalent is that ψ is zero for chains of GETs such as the SHIFT network (Fig. 5.3a). This enables ψ to better match our intuition for "cooperation among diverse parts". For example, in Figure 5.3 the network with the highest ϕ is the counter-intuitive SHIFT; on the other hand, the network with the highest ψ is the more sensible 4321 (see bottom table in Figure 5.4).

The third row in Figure 5.4 shows a difference related to KEEPs vs GETs—how ψ and ϕ respectively treat self-connections. In ANDtriplet (Figure 5.4e) each node integrates information about two nodes. Likewise, in iso-ANDtriplet (Figure 5.4f) each node integrates information about two nodes, but the information is about *itself* and one other.

Just as ψ views KEEP and GET nodes equivalently, ψ views self and cross connections equivalently. In fact, by property (**Eq**) the ψ values for ANDtriplet and iso-ANDtriplet are provably equal. On the other hand, ϕ considers self and cross connections differently in that ϕ can only decrease when adding a self-connection. As such, the ϕ for iso-ANDtriplet is less than ANDtriplet.

The fourth row in Figure 5.4 shows this same self-connections business carrying over to duplicate computations. Although AND-AND (Figure 5.4h) and AND-ZERO (Figure 5.4g) perform the same computation, AND-AND has an additional self-connection that pushes AND-AND's ϕ below that of AND-ZERO. By (**Eq**), ψ is provably invariant under such duplicate computations.

5.7 Conclusion

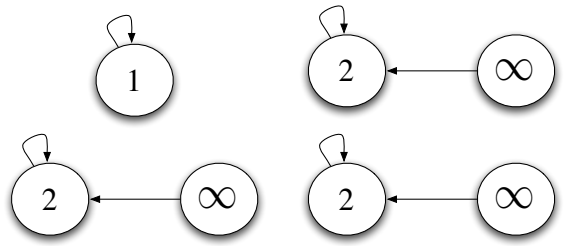
Regardless of any connection to consciousness, and purely as a measure of functional irreducibility, we have three concerns about ϕ: (1) state-dependent ϕ can exceed the entropy of the entire system; (2) ϕ often decreases with duplicate computation; (3) ϕ doesn't match the intuition of "cooperation among diverse parts".

We introduced a new irreducibility measure, ψ, that solves all three concerns and otherwise stays close to the original spirit of ϕ—i.e., the quantification of a system's irreducibility to disjoint parts. Based in Partial Information Decomposition, ψ has other desirable properties, such as not needing a MIP normalization and being substantially faster to compute.

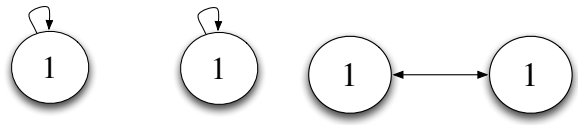
Finally, we contrasted ψ versus ϕ with simple, concrete examples.

Although we recommend using ψ over ϕ, the ψ measure remains imperfect. The most notable areas for improvement are:

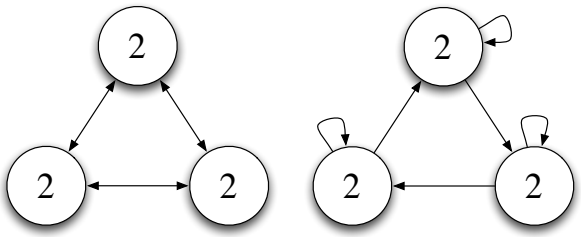
1. The current ψ bounds are too loose. We need to tighten the $\mathrm{I}_\cap$ bounds, which will tighten the



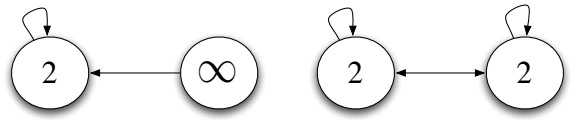
(a) AND-ZERO+KEEP (b) 2x AND-ZERO



(c) KEEP-KEEP (d) GET-GET



(e) ANDtriplet (f) iso-ANDtriplet



(g) AND-ZERO (h) AND-AND

Network	I($X:Y$)	$\langle\phi\rangle$	$\langle\psi\rangle_{\min}$	$\langle\psi\rangle_{\max}$
AND-ZERO+KEEP (a)	1.81	0	0	0.50
2x AND-ZERO (b)	1.62	0	0	0.50
KEEP-KEEP (c)	2.00	0	0	1.00
GET-GET (d)	2.00	2.00	0	1.00
SHIFT (Fig. 5.3a)	4.00	2.00	0	1.00
4422 (Fig. 5.3b)	1.20	0.42	0.33	0.50
4322 (Fig. 5.3c)	1.81	1.37	0.68	0.88
4321 (Fig. 5.3d)	2.03	1.65	0.78	1.00
ANDtriplet (e)	2.00	2.00	0.16	0.75
iso-ANDtriplet (f)	2.00	1.07	0.16	0.75
AND-ZERO (g)	0.81	0.50	0.19	0.5
AND-AND (h)	0.81	0.19	0.19	0.5

Figure 5.4: Contrasting $\langle\phi\rangle$ versus $\langle\psi\rangle$ for exemplary networks.

derived bounds on ψ and $\langle\psi\rangle$.

2. Justify why a measure of conscious experience should privilege irreducibility to disjoint parts over irreducibility to overlapping parts.

3. Reformalize the work on qualia in [4] using ψ.

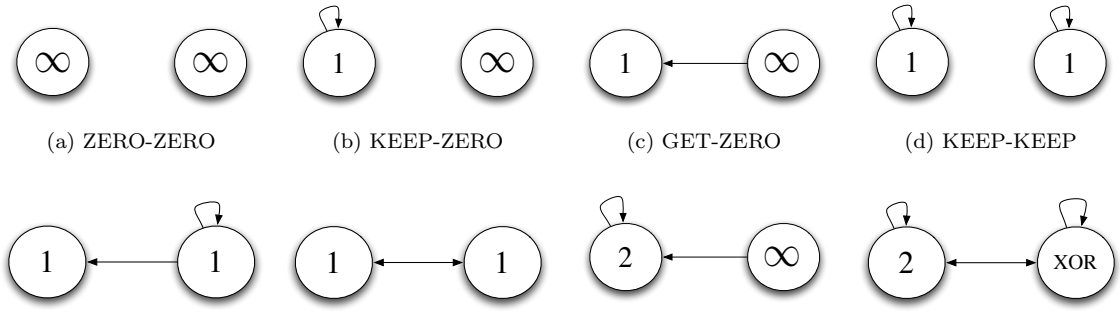
4. Although not specific to ψ, there needs to be a stronger justification for the chosen distribution on X.

Appendix

5.A Reading the network diagrams

We present eight doublet networks and their transition tables so you can see how the network diagram specifies the transition table. Figure 5.5 shows eight network diagrams to build your intuition. The number inside each node is that node's *activation threshold*. A node updates to `1` (conceptually an "ON") if there are at least as many of inputs ON as its activation threshold; e.g. a node with an inscribed 2 updates to a `1` if two or more incoming wires are ON. An activation threshold of ∞ means the node always updates to `0` (conceptually an "OFF"). A binary string denotes the state of the network, read left to right.

We take the AND-ZERO network (Figure 5.5g) as an example. Although the AND-ZERO network can never output `01` or `11` (Fig. 1b), we still consider states `01`, `11` as equally possible states at time=0. This is because X_0 is uniformly distributed per assumption **(A)**. The state of the AND-node (left) at time=1 is a function of *both* nodes at time=0. For example, in the AND-ZERO gate, the left binary digit is the state of the AND-node and the right binary digit is the state of the ZERO-node.

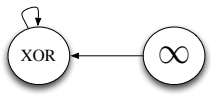


(a) ZERO-ZERO (b) KEEP-ZERO (c) GET-ZERO (d) KEEP-KEEP

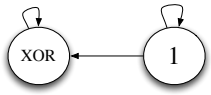
(e) GET-KEEP (f) GET-GET (g) AND-ZERO (h) AND-XOR

X		ZERO-ZERO	KEEP-ZERO	GET-ZERO	KEEP-KEEP	GET-KEEP	GET-GET	AND-ZERO	AND-XOR
00	→	00	00	00	00	00	00	00	00
01	→	00	00	10	01	11	10	00	01
10	→	00	10	00	10	00	01	00	01
11	→	00	10	10	11	11	11	10	10

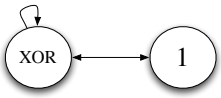
Figure 5.5: Eight doublet networks with transition tables.



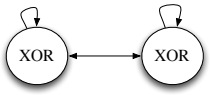
(a) XOR-ZERO



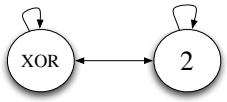
(b) XOR-KEEP



(c) XOR-GET



(d) XOR-XOR



(e) XOR-AND

X		XOR-ZERO	XOR-KEEP	XOR-GET	XOR-XOR	XOR-AND
00	→	00	00	00	00	00
01	→	10	11	10	11	10
10	→	10	10	11	11	10
11	→	00	01	01	00	01

Network	I($X:Y$)	$\langle\phi\rangle$	$\langle\psi\rangle_{\min}$	$\langle\psi\rangle_{\max}$
ZERO-ZERO (Fig. 5.5a)	0	0	0	0
KEEP-ZERO (Fig. 5.5b)	1.0	0	0	0
KEEP-KEEP (Fig. 5.5d)	2.0	0	0	1.0
GET-ZERO (Fig. 5.5c)	1.0	1.0	0	0
GET-KEEP (Fig. 5.5e)	1.0	0	0	0
GET-GET (Fig. 5.5f)	2.0	2.0	0	1.0
AND-ZERO (Fig. 5.2a)	0.811	0.5	0.189	0.5
AND-KEEP	1.5	0.189	0	0.5
AND-GET	1.5	1.189	0	0.5
AND-AND (Fig. 5.2b)	0.811	0.189	0.189	0.5
AND-XOR (Fig. 5.5h)	1.5	1.189	0.5	1.0
XOR-ZERO (a)	1.0	1.0	1.0	1.0
XOR-KEEP (b)	2.0	1.0	0	1.0
XOR-GET (c)	2.0	2.0	0	1.0
XOR-AND (e)	1.5	1.189	0.5	1.0
XOR-XOR (d)	1.0	1.0	1.0	1.0

Figure 5.6: Networks, transition tables, and measures for the diagnostic doublets.

5.B Necessary proofs

5.B.1 Proof that the max union of bipartions covers all partitions

Lemma 5.B.1. *Given properties* $(\mathbf{S_0})$ *and* $(\mathbf{M_0})$*, the maximum union information conveyed by a partition of predictors* $\mathbf{X} = \{X_1, \ldots, X_n\}$ *about state* $Y = y$ *equals the maximum union information conveyed by a bipartition of* $\mathbf{X}$ *about state* $Y = y$*.*

Proof. We prove that the maximum Information conveyed by a Partition, $\mathsf{IbDp}(\mathbf{X} : y)$, equals Information conveyed by a Bipartition, $\mathsf{IbB}(\mathbf{X} : y)$ by showing,

$$\mathsf{IbDp}(\mathbf{X} : y) \leq \mathsf{IbB}(\mathbf{X} : y) \leq \mathsf{IbDp}(\mathbf{X} : y) \ . \tag{5.11}$$

We first show that $\mathsf{IbB}(\mathbf{X} : y) \leq \mathsf{IbDp}(\mathbf{X} : y)$. By their definitions,

$$\begin{aligned} \mathsf{IbDp}(\mathbf{X} : y) &\equiv \max_{\mathbf{P}} \mathrm{I}_{\cup}(\mathbf{P} : y) \\ \mathsf{IbB}(\mathbf{X} : y) &\equiv \max_{\substack{\mathbf{P} \\ |\mathbf{P}|=2}} \mathrm{I}_{\cup}(\mathbf{P} : y) \ , \end{aligned} \tag{5.12}$$

where $\mathbf{P}$ enumerates over all partitions of set $\mathbf{X}$.

By removing the restriction that $|\mathbf{P}| = 2$ from the maximization in IbB we arrive at IbDp. As removing a restriction can only increase the maximum, thus $\mathsf{IbB}(\mathbf{X} : y) \leq \mathsf{IbDp}(\mathbf{X} : y)$.

We next show that $\mathsf{IbDp}(\mathbf{X} : y) \leq \mathsf{IbB}\,(\mathbf{X} : y)$, meaning we must show that,

$$\max_{\mathbf{P}} \ \mathrm{I}_{\cup}(\mathbf{P} : y) \leq \max_{\substack{\mathbf{P} \\ |\mathbf{P}|=2}} \mathrm{I}_{\cup}(\mathbf{P} : y) \ . \tag{5.13}$$

Without loss of generality, we choose an arbitrary subset/part $S \subset \mathbf{X}$. This yields the bipartition of parts $\{S, \mathbf{X} \setminus S\}$. We then further partition the second part, $\mathbf{X} \setminus S$, into k disjoint subcomponents denoted $\{T_1, \ldots, T_k\}$ where $2 \leq k \leq n - |S|$, creating an arbitrary partition $\mathbf{P} = \{S, T_1, \ldots, T_k\}$. We now need to show that,

$$\mathrm{I}_{\cup}(S, T_1, \ldots, T_k : y) \leq \mathrm{I}_{\cup}\big(S, \mathbf{X} \setminus S : y\big) \ .$$

By $(\mathbf{M_0})$ equality condition, we can append each subcomponent $T_1, \ldots, T_k$ to $\{S, \mathbf{X} \setminus S\}$ without changing the union-information, because every subcomponent $T_i \preceq \mathbf{X} \setminus S$. Then applying $(\mathbf{S_0})$, we re-order the parts so that $S, T_1, \ldots, T_k$ come first. This yields,

$$\mathrm{I}_{\cup}(S, T_1, \ldots, T_k : y) \leq \mathrm{I}_{\cup}\big(S, T_1, \ldots, T_k, \mathbf{X} \setminus S : y\big) \ .$$

Applying ($\mathbf{M_0}$) inequality condition, adding the predictor $\mathbf{X} \setminus S$ can only increase the union information. Therefore we prove eq. (5.13), which proves eq. (5.11), that $\mathsf{IbDp}(\mathbf{X} : y) = \mathsf{IbB}(\mathbf{X} : y)$. $\square$

5.B.2 Bounds on $\psi(X_1, \ldots, X_n : y)$

Lemma 5.B.2. *Given* $(\mathbf{M_0})$, $(\mathbf{SR})$ *and the predictors* $X_1, \ldots, X_n$ *are independent, i.e.* $\mathrm{H}(X_{1\ldots n}) = \sum_{i=1}^{n} \mathrm{H}(X_i)$*, then,*

$$\psi(X_1, \ldots, X_n : y) \leq \min_{i \in \{1,\ldots,n\}} \mathrm{D_{KL}} \left[\Pr(X_{1\ldots n}|y) \middle\| \Pr(X_i) \Pr\Big(X_{1\ldots n\setminus i} \Big| y\Big) \right] \ .$$

Proof. Applying $(\mathbf{M_0})$ inequality condition, we have $\mathrm{I}_\cap(A, B : y) \leq \min\big[\mathrm{I}(A:y), \mathrm{I}(B:y)\big]$. Via the inclusion-exclusion rule, this entails $\mathrm{I}_\cup(A, B : y) \geq \max\big[\mathrm{I}(A:y), \mathrm{I}(B:y)\big]$, and we use this to upper-bound $\psi(X_1, \ldots, X_n : y)$. The random variable $A \neq \emptyset$, $B \equiv \mathbf{X} \setminus A$, and $AB \equiv X_{1\ldots n}$.

$$\begin{aligned}
\psi(X_1, \ldots, X_n : y) &= \mathrm{I}(X_{1\ldots n} : y) - \max_{A \subset \mathbf{X}} \mathrm{I}_\cup(A, B : y) \\
&\leq \mathrm{I}(X_{1\ldots n} : y) - \max_{A \subset \mathbf{X}} \max\big[\mathrm{I}(A:y), \mathrm{I}(B:y)\big] \\
&\quad \text{By symmetry of complementary bipartitions, every } B \text{ will be an } A \text{ at some} \\
&\quad \text{point. So we can drop the } B \text{ term.} \\
&= \mathrm{I}(X_{1\ldots n} : y) - \max_{A \subset \mathbf{X}} \mathrm{I}(A:y) \ .
\end{aligned}$$

For two random variables A and A' such that $A \preceq A'$, $\mathrm{I}(A:y) \leq \mathrm{I}\big(A':y\big)$.[13] Therefore, there will always be a maximizing subset of $\mathbf{X}$ with size $n-1$.

$$\begin{aligned}
\psi(X_1, \ldots, X_n : y) &\leq \mathrm{I}(X_{1\ldots n} : y) - \max_{\substack{A \subset \mathbf{X} \\ |A| = n-1}} \mathrm{I}(A:y) \\
&= \mathrm{I}(X_{1\ldots n} : y) - \max_{i \in \{1,\ldots,n\}} \mathrm{I}\Big(X_{1\ldots n\setminus i} : y\Big) \\
&= \min_{i \in \{1,\ldots,n\}} \mathrm{I}(X_{1\ldots n} : y) - \mathrm{I}\Big(X_{1\ldots n\setminus i} : y\Big) \\
&= \min_{i \in \{1,\ldots,n\}} \mathrm{I}\Big(X_i : y \Big| X_{1\ldots n\setminus i}\Big) \\
&= \min_{i \in \{1,\ldots,n\}} \mathrm{D_{KL}} \left[\Pr(X_{1\ldots n}|y) \middle\| \Pr\Big(X_i \Big| X_{1\ldots n\setminus i}\Big) \Pr\Big(X_{1\ldots n\setminus i} \Big| y\Big) \right] \ .
\end{aligned}$$

Now applying that the predictors $\mathbf{X}$ are independent, $\Pr\Big(x_i \Big| x_{1\ldots n\setminus i}\Big) = \Pr(x_i)$. This yields,

$$\psi(X_1, \ldots, X_n : y) \leq \min_{i \in \{1,\ldots,n\}} \mathrm{D_{KL}} \left[\Pr(X_{1\ldots n}|y) \middle\| \Pr(X_i) \Pr\Big(X_{1\ldots n\setminus i} \Big| y\Big) \right] \ .$$

[13] $\mathrm{I}(A:y) \leq \mathrm{I}\big(A':y\big)$ because $\mathrm{I}\big(A':y\big) = \mathrm{I}(A:y) + \mathrm{I}\big(A':y|A\big)$.

□

Lemma 5.B.3. *Given* (**GP**), (**SR**) *and predictors* $X_1, \ldots, X_n$ *are independent, i.e.* $\mathrm{H}(X_{1\ldots n}) = \sum_{i=1}^{n} \mathrm{H}(X_i)$*, then,*

$$\begin{aligned}\psi(X_1, \ldots, X_n : y) &\geq \min_{A \subset \mathbf{X}} \mathrm{I}(A : B|y) \\ &= \min_{A \subset \mathbf{X}} \mathrm{D}_{\mathrm{KL}} \left[\Pr\big(X_{1\ldots n}|y\big) \middle\| \Pr\big(A|y\big) \Pr\big(B|y\big) \right] .\end{aligned}$$

Proof. First, from the definition of $\mathrm{I}_\cup$, $\mathrm{I}_\cup(A, B : y) = \mathrm{I}(A : y) + \mathrm{I}(B : y) - \mathrm{I}_\cap(A, B : y)$. Then applying (**GP**), we have $\mathrm{I}_\cup(A, B : y) \leq \mathrm{I}(A : y) + \mathrm{I}(B : y)$. We use this to lowerbound $\psi(X_1, \ldots, X_n : y)$. The random variable $A \neq \emptyset$, $B \equiv \mathbf{X} \setminus A$, and $AB \equiv X_{1\ldots n}$.

$$\begin{aligned}\psi(X_1, \ldots, X_n : y) &= \mathrm{I}(X_{1\ldots n} : y) - \max_{A \subset \mathbf{X}} \mathrm{I}_\cup(A, B : y) \\ &\geq \mathrm{I}(X_{1\ldots n} : y) - \max_{A \subset \mathbf{X}} \big[\mathrm{I}(A : y) + \mathrm{I}(B : y) \big] \\ &= \min_{A \subset \mathbf{X}} \mathrm{I}(AB : y) - \mathrm{I}(A : y) - \mathrm{I}(B : y) \\ &= \min_{A \subset \mathbf{X}} \mathrm{I}\big(A : y|B\big) - \mathrm{I}(A : y) \\ &= \min_{A \subset \mathbf{X}} \mathrm{D}_{\mathrm{KL}} \left[\Pr(AB|y) \middle\| \Pr\big(B|y\big) \Pr\big(A|B\big) \right] - \mathrm{D}_{\mathrm{KL}} \big[\Pr(A|y) \big\| \Pr(A) \big] \\ &= \min_{A \subset \mathbf{X}} \sum_{a,b} \Pr\big(ab|y\big) \log \frac{\Pr\big(ab|y\big)}{\Pr\big(b|y\big) \Pr\big(a|b\big)} + \sum_{a} \Pr\big(a|y\big) \log \frac{\Pr(a)}{\Pr\big(a|y\big)} .\end{aligned}$$

We now add $\sum_b \Pr(b|ay)$ in front of the right-most $\sum_a$. We can do this because $\sum_b \Pr(b|ay) = 1.0$. This then yields,

$$\begin{aligned}\psi(X_1, \ldots, X_n : y) &\geq \min_{A \subset \mathbf{X}} \sum_{a,b} \Pr\big(ab|y\big) \log \frac{\Pr\big(ab|y\big)}{\Pr\big(b|y\big) \Pr\big(a|b\big)} + \Pr(b|ay) \Pr\big(a|y\big) \log \frac{\Pr(a)}{\Pr\big(a|y\big)} \\ &= \min_{A \subset \mathbf{X}} \sum_{a,b} \Pr\big(ab|y\big) \left[\log \frac{\Pr\big(ab|y\big)}{\Pr\big(b|y\big) \Pr\big(a|b\big)} + \log \frac{\Pr(a)}{\Pr\big(a|y\big)} \right] \\ &= \min_{A \subset \mathbf{X}} \sum_{a,b} \Pr(ab|y) \log \frac{\Pr\big(ab|y\big) \Pr(a)}{\Pr\big(a|y\big) \Pr\big(b|y\big) \Pr\big(a|b\big)} .\end{aligned}$$

Now applying that the predictors $\mathbf{X}$ are independent, $\Pr\big(a|b\big) = \Pr(a)$; thus we can cancel $\Pr(a)$

for $\Pr\big(a|b\big)$. This yields,

$$\begin{aligned}\psi(X_1,\ldots,X_n:y) &\geq \min_{A\subset\mathbf{X}} \sum_{a,b} \Pr(ab|y)\log\frac{\Pr\big(ab|y\big)}{\Pr\big(a|y\big)\Pr\big(b|y\big)} \\ &= \min_{A\subset\mathbf{X}} \mathrm{D}_{\mathrm{KL}}\left[\Pr\big(X_{1\ldots n}|y\big)\middle\|\Pr\big(A|y\big)\Pr\big(B|y\big)\right] \ .\end{aligned}$$

□

5.B.3 Bounds on $\langle\psi\rangle(X_1,\ldots,X_n:Y)$

Lemma 5.B.4. *Given* (**M₀**), (**SR**) *and the predictors* $X_1,\ldots,X_n$ *are independent, i.e.* $\mathrm{H}(X_{1\ldots n}) = \sum_{i=1}^{n}\mathrm{H}(X_i)$*, then,*

$$\langle\psi\rangle(X_1,\ldots,X_n:Y) \leq \min_{i\in\{1,\ldots,n\}} \mathrm{D_{KL}}\left[\Pr(X_{1\ldots n},Y)\middle\|\Pr\Big(X_{1\ldots n\setminus i},Y\Big)\Pr(X_i)\right] .$$

Proof. First, using the same reasoning in Lemma 5.B.2, we have,

$$\begin{aligned}
\langle\psi\rangle(\mathbf{X}:Y) &\leq \mathrm{I}(X_{1\ldots n}{:}Y) - \max_{i\in\{1,\ldots,n\}} \mathrm{I}\Big(X_{1\ldots n\setminus i}{:}Y\Big)\\
&= \min_{i\in\{1,\ldots,n\}} \mathrm{I}(X_{1\ldots n}{:}Y) - \mathrm{I}\Big(X_{1\ldots n\setminus i}{:}Y\Big)\\
&= \min_{i\in\{1,\ldots,n\}} \mathrm{I}\Big(X_i{:}Y\Big|X_{1\ldots n\setminus i}\Big)\\
&= \min_{i\in\{1,\ldots,n\}} \mathrm{D_{KL}}\left[\Pr(X_{1\ldots n},Y)\middle\|\Pr\Big(X_i\Big|X_{1\ldots n\setminus i}\Big)\Pr\Big(X_{1\ldots n\setminus i},Y\Big)\right] .
\end{aligned}$$

Now applying that the predictors $\mathbf{X}$ are independent, $\Pr\Big(X_i\Big|X_{1\ldots n\setminus i}\Big) = \Pr(X_i)$. This yields,

$$\langle\psi\rangle(\mathbf{X}:Y) \leq \min_{i\in\{1,\ldots,n\}} \mathrm{D_{KL}}\left[\Pr(X_{1\ldots n},Y)\middle\|\Pr\Big(X_{1\ldots n\setminus i},Y\Big)\Pr(X_i)\right] .$$

□

Lemma 5.B.5. *Given* (**GP**), (**SR**) *and predictors* $X_1,\ldots,X_n$ *are independent, i.e.* $\mathrm{H}(X_{1\ldots n}) = \sum_{i=1}^{n}\mathrm{H}(X_i)$*, then,*

$$\langle\psi\rangle(X_1,\ldots,X_n:Y) \geq \min_{A\subset\mathbf{X}} \mathrm{I}(A{:}B|Y) .$$

Proof. First, using the same reasoning in Lemma 5.B.3, we have,

$$\begin{aligned}
\langle\psi\rangle(X_1,\ldots,X_n:Y) &\geq \mathrm{I}(X_{1\ldots n}{:}Y) - \max_{A\subset\mathbf{X}}\big[\mathrm{I}(A{:}Y)+\mathrm{I}(B{:}Y)\big]\\
&= \min_{A\subset\mathbf{X}} \mathrm{I}(AB{:}Y) - \mathrm{I}(A{:}Y) - \mathrm{I}(B{:}Y)\\
&= \min_{A\subset\mathbf{X}} \mathrm{I}(A{:}B|Y) - \mathrm{I}(A{:}B) .
\end{aligned}$$

Now applying that the predictors $\mathbf{X}$ are independent, $\mathrm{I}(A\!:\!B) = 0$. This yields,

$$\langle\psi\rangle(X_1, \ldots, X_n : Y) \geq \min_{A \subset \mathbf{X}} \mathrm{I}(A\!:\!B|Y) \ .$$

□

5.C Definition of intrinsic $\texttt{ei}(y/\mathbf{P})$ a.k.a. "perturbing the wires"

State-dependent `ei` across a partition, fully written as $\texttt{ei}\big(X \to y/\mathbf{P}\big)$ and abbreviated $\texttt{ei}\big(y/\mathbf{P}\big)$, is defined by eq. (5.14). The probability distribution of the "intrinsic information" in the entire system, $\Pr(X \to y)$, is simply $\Pr\big(X|y\big)$ (eq. (5.15)).[14]

$$\begin{aligned} \texttt{ei}\big(X \to y/\mathbf{P}\big) &\equiv \mathrm{D}_{\mathrm{KL}}\left[\Pr^*(X \to y)\middle\|\prod_{i=1}^{m}\Pr\Big(X_i^{\mathbf{P}} \to y_i^{\mathbf{P}}\Big)\right] && (5.14)\\ &= \mathrm{D}_{\mathrm{KL}}\left[\Pr\big(X|y\big)\middle\|\prod_{i=1}^{m}\Pr^*\big(X_i^{\mathbf{P}}\big|y_i^{\mathbf{P}}\big)\right] . && (5.15)\end{aligned}$$

Balduzzi/Tononi [3] define the probability distribution describing the intrinsic information from the whole system X to state y as,

$$\Pr(X \to y) = \Pr\big(X|y\big) = \Big\{\Pr\big(x|y\big) : \forall x \in X\Big\} .$$

They then define probability distribution describing the intrinsic information from a part $X_i^{\mathbf{P}}$ to a state $y_i^{\mathbf{P}}$ as,

$$\Pr^*\big(X_i^{\mathbf{P}} \to y_i^{\mathbf{P}}\big) \equiv \Pr^*\big(X_i^{\mathbf{P}}\big|Y_i^{\mathbf{P}} = y_i^{\mathbf{P}}\big) = \Big\{\Pr^*\big(x_i^{\mathbf{P}}\big|y_i^{\mathbf{P}}\big) : \forall x_i^{\mathbf{P}} \in X_i^{\mathbf{P}}\Big\} .$$

First we define the fundamental property of the $\Pr^*$ distribution. Given a state $x_i^{\mathbf{P}}$, the probability of a state $y_i^{\mathbf{P}}$ is computed by probability that each node in the state $y_i^{\mathbf{P}}$ independently reaches the state specified by $y_i^{\mathbf{P}}$,

$$\Pr^*\big(y_i^{\mathbf{P}}\big|x_i^{\mathbf{P}}\big) \equiv \prod_{j=1}^{|\mathbf{P}_i|}\Pr\Big(y_{i,j}^{\mathbf{P}}\Big|x_i^{\mathbf{P}}\Big) . \tag{5.16}$$

Then we define the join distribution relative to eq. (5.16):

$$\begin{aligned}\Pr^*\big(x_i^{\mathbf{P}}, y_i^{\mathbf{P}}\big) &= \Pr^*\big(x_i^{\mathbf{P}}\big)\Pr^*\big(y_i^{\mathbf{P}}\big|x_i^{\mathbf{P}}\big)\\ &= \Pr^*\big(x_i^{\mathbf{P}}\big)\textstyle\prod_{j=1}^{|\mathbf{P}_i|}\Pr\Big(y_{i,j}^{\mathbf{P}}\Big|x_i^{\mathbf{P}}\Big) .\end{aligned}$$

Then applying assumption **(B)**, X follows a discrete uniform distribution, so $\Pr^*\big(x_i^{\mathbf{P}}\big) \equiv \Pr\big(x_i^{\mathbf{P}}\big) =$

[14] It's worth nothing that $\Pr^*\big(X|y\big) \neq \Pr\big(X|y\big)$.

$1/|X_i^{\mathbf{P}}|$. This gives us the complete definition of $\Pr^*\big(x_i^{\mathbf{P}}, y_i^{\mathbf{P}}\big)$,

$$\Pr^*\big(x_i^{\mathbf{P}}, y_i^{\mathbf{P}}\big) = \Pr\big(x_i^{\mathbf{P}}\big) \prod_{j=1}^{|\mathbf{P}_i|} \Pr\Big(y_{i,j}^{\mathbf{P}} \Big| x_i^{\mathbf{P}}\Big) \ . \tag{5.17}$$

With the joint $\Pr^*$ distribution defined, we can compute anything we want by summing over the eq. (5.17)—such as the expressions for $\Pr^*\big(y_i^P\big)$ and $\Pr^*\big(x_i^{\mathbf{P}} \big| y_i^{\mathbf{P}}\big)$,

$$\begin{aligned} \Pr^*\big(y_i^{\mathbf{P}}\big) &= \sum_{x_i^{\mathbf{P}} \in X_i^{\mathbf{P}}} \Pr^*\big(x_i^{\mathbf{P}}, y_i^{\mathbf{P}}\big) \\ \Pr^*\big(x_i^{\mathbf{P}} \big| y_i^{\mathbf{P}}\big) &= \frac{\Pr^*\big(x_i^{\mathbf{P}}, y_i^{\mathbf{P}}\big)}{\Pr^*\big(y_i^{\mathbf{P}}\big)} \ . \end{aligned}$$

5.D Misc proofs

Given the properties (**GP**), (**SR**), and the predictors $X_1, \ldots, X_n$ are independent, i.e. $\mathrm{H}(X_{1\ldots n}) = \sum_{i=1}^{n} \mathrm{H}(X_i)$, we show that $\langle\phi\rangle \not\leq \langle\psi\rangle$. This is equivalent to,

$$\langle\psi\rangle_{\min}(\mathbf{X} : Y) \leq \langle\phi\rangle(\mathbf{X} : Y) \ .$$

Proof. We prove the above by showing that for any bipartition $\mathbf{P}$, that $\langle\psi\rangle_{\min}(\mathbf{X} : Y) \leq \big\langle \mathtt{ei}(Y/\mathbf{P})\big\rangle$.

For a bipartition $\mathbf{P}$,

$$\begin{aligned}
\langle\psi\rangle_{\min}(\mathbf{X} : Y) &= \mathrm{I}\Big(X_1^{\mathbf{P}} : X_2^{\mathbf{P}} \Big| Y\Big) \\
&= \mathrm{I}\Big(X_1^{\mathbf{P}} : X_2^{\mathbf{P}} \Big| Y\Big) - \mathrm{I}\Big(X_1^{\mathbf{P}} : X_2^{\mathbf{P}}\Big) \\
&= \mathrm{I}(X : Y) - \mathrm{I}\Big(X_1^{\mathbf{P}} : Y\Big) - \mathrm{I}\Big(X_2^{\mathbf{P}} : Y\Big) \\
\langle \mathtt{ei}(Y/\mathbf{P})\rangle &= \mathrm{I}(X : Y) - \mathrm{I}\Big(X_1^{\mathbf{P}} : Y_1^{\mathbf{P}}\Big) - \mathrm{I}\Big(X_1^{\mathbf{P}} : Y_1^{\mathbf{P}}\Big) \ .
\end{aligned}$$

$$\begin{aligned}
\langle \mathtt{ei}(Y/\mathbf{P})\rangle - \langle\psi\rangle_{\min}(\mathbf{X} : Y) &= \mathrm{I}(X : Y) - \mathrm{I}\Big(X_1^{\mathbf{P}} : Y_1^{\mathbf{P}}\Big) - \mathrm{I}\Big(X_1^{\mathbf{P}} : Y_1^{\mathbf{P}}\Big) - \mathrm{I}(X : Y) + \mathrm{I}\Big(X_1^{\mathbf{P}} : Y\Big) + \mathrm{I}\Big(X_2^{\mathbf{P}} : Y\Big) \\
&= \mathrm{I}\Big(X_1^{\mathbf{P}} : Y\Big) - \mathrm{I}\Big(X_1^{\mathbf{P}} : Y_1^{\mathbf{P}}\Big) + \mathrm{I}\Big(X_2^{\mathbf{P}} : Y\Big) - \mathrm{I}\Big(X_2^{\mathbf{P}} : Y_2^{\mathbf{P}}\Big) \\
&= \mathrm{I}\Big(X_1^{\mathbf{P}} : Y_2^{\mathbf{P}} \Big| Y_1^{\mathbf{P}}\Big) + \mathrm{I}\Big(X_2^{\mathbf{P}} : Y_1^{\mathbf{P}} \Big| Y_2^{\mathbf{P}}\Big) \\
&\geq 0 \ .
\end{aligned}$$

And we complete the proof that $\langle\psi\rangle_{\min} \leq \langle\phi\rangle$. Therefore, $\langle\phi\rangle \not\leq \langle\psi\rangle$.

□

5.E Setting $t = 1$ without loss of generality

Given t stationary surjective functions that may be different or the same, denoted $f_1 \cdots f_t$, we define the state of system at time t, denoted X_t, as the application of the t functions to the state of the system at time 0, denoted X_0,

$$X_t = f_t \left(f_{t-1} \left(\cdots f_2 \left(f_1 \left(X_0 \right) \right) \cdots \right) \right) .$$

We instantiate an empty "dictionary function" $g(\bullet)$. Then for every $x_0 \in X_0$ we assign,

$$g(x_0) \equiv f_t \left(f_{t-1} \left(\cdots f_2 \left(f_1 \left(x_0 \right) \right) \cdots \right) \right) . \qquad \forall x_0 \in X_0$$

At the end of this process we have a function g that accomplishes any chain of stationary functions $f_1 \cdots f_t$ in a single step for the entire domain of f_1. So instead of studying the transformation,

$$X_0 \xrightarrow{f_1 \cdots f_t} X_t ,$$

we can equivalently study the transformation,

$$X_0 \xrightarrow{g} Y .$$

Here's an example using mechanism $f_1 = f_2 = f_3 = f_4 =$ AND-GET.

time=0		$t=1$		$t=2$		$t=3$		$t=4$
00	$\rightarrow$	00	$\rightarrow$	00	$\rightarrow$	00	$\rightarrow$	00
01	$\rightarrow$	00	$\rightarrow$	00	$\rightarrow$	00	$\rightarrow$	00
10	$\rightarrow$	01	$\rightarrow$	00	$\rightarrow$	00	$\rightarrow$	00
11	$\rightarrow$	11	$\rightarrow$	11	$\rightarrow$	10	$\rightarrow$	00
$g(\bullet)$		AND-GET		AND-AND		AND-ZERO		ZERO-ZERO

Table 5.1: Applying the update rule "AND-GET", over four timesteps.

Bibliography

[1] Dimitris Anastassiou. Computational analysis of the synergy among multiple interacting genes. *Molecular Systems Biology*, 3:83, 2007.

[2] David Balduzzi. personal communication.

[3] David Balduzzi and Giulio Tononi. Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Computational Biology*, 4(6):e1000091, Jun 2008.

[4] David Balduzzi and Giulio Tononi. Qualia: The geometry of integrated information. *PLoS Computational Biology*, 5(8), 2009.

[5] Adam B. Barett and Anil K. Seth. Practical measures of integrated information for time-series data. *PLoS Computational Biology*, 2010.

[6] Anthony J. Bell. The co-information lattice. In S. Amari, A. Cichocki, S. Makino, and N. Murata, editors, *Fifth International Workshop on Independent Component Analysis and Blind Signal Separation*. Springer, 2003.

[7] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, and Jürgen Jost. Shared information – new insights and problems in decomposing information in complex systems. *CoRR*, abs/1210.5902, 2012.

[8] N. J. Cerf and C. Adami. Negative entropy and information in quantum mechanics. *Phys. Rev. Lett.*, 79:5194–5197, Dec 1997.

[9] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley, New York, NY, 1991.

[10] M. R. DeWeese and M. Meister. How to measure the information gained from one symbol. *Network*, 10:325–340, Nov 1999.

[11] T. G. Dietterich, S. Becker, and Z. Ghahramani, editors. *Group Redundancy Measures Reveal Redundancy Reduction in the Auditory Pathway*, Cambridge, MA, 2002. MIT Press.

[12] Itay Gat and Naftali Tishby. Synergy and redundancy among brain cells of behaving monkeys. In *Advances in Neural Information Proceedings systems*, pages 465–471. MIT Press, 1999.

[13] Timothy J. Gawne and Barry J. Richmond. How independent are the messages carried by adjacent inferior temporal cortical neurons? *Journal of Neuroscience*, 13:2758–71, 1993.

[14] V. Griffith, E. K. P. Chong, R. G. James, C. J. Ellison, and J. P. Crutchfield. Intersection information based on common randomness. *ArXiv e-prints*, October 2013.

[15] Virgil Griffith and Christof Koch. Quantifying synergistic mutual information. In M. Prokopenko, editor, *Guided Self-Organization: Inception*. Springer, 2014.

[16] Peter Gács and Jack Körner. Common information is far less than mutual information. *Problems of Control and Informaton Theory*, 2(2):149–162, 1973.

[17] Te Sun Han. Nonnegative entropy measures of multivariate symmetric correlations. *Information and Control*, 36(2):133–156, 1978.

[18] Malte Harder, Christoph Salge, and Daniel Polani. A bivariate measure of redundant information. *CoRR*, abs/1207.2080, 2012.

[19] A. Jakulin and I. Bratko. Analyzing attribute dependencies. In *Lecture Notes in Artificial Intelligence*, pages 229–240, 2838 2003.

[20] Kevin B Korb, Lucas R Hope, and Erik P Nyberg. Information-theoretic causal power. In *Information Theory and Statistical Learning*, pages 231–265. Springer, 2009.

[21] Peter E. Latham and Sheila Nirenberg. Synergy, redundancy, and independence in population codes, revisited. *Journal of Neuroscience*, 25(21):5195–5206, May 2005.

[22] Hua Li and Edwin K. P. Chong. On a connection between information and group lattices. *Entropy*, 13(3):683–708, 2011.

[23] Joseph T. Lizier, Benjamin Flecker, and Paul L. Williams. Towards a synergy-based approach to measuring information modification. *CoRR*, abs/1303.3440, 2013.

[24] W. J. McGill. Multivariate information transmission. *Psychometrika*, 19:97–116, 1954.

[25] S Nirenberg, S M Carcieri, A L Jacobs, and P E Latham. Retinal ganglion cells act largely as independent encoders. *Nature*, 411(6838):698–701, Jun 2001.

[26] Sheila Nirenberg and Peter E. Latham. Decoding neuronal spike trains: How important are correlations? *Proceedings of the National Academy of Sciences*, 100(12):7348–7353, 2003.

[27] S Panzeri, A Treves, S Schultz, and E T Rolls. On decoding the responses of a population of neurons from short time windows. *Neural Comput*, 11(7):1553–1577, Oct 1999.

[28] G Pola, A Thiele, K P Hoffmann, and S Panzeri. An exact method to quantify the information transmitted by different mechanisms of correlational coding. *Network*, 14(1):35–60, Feb 2003.

[29] E. Schneidman, W. Bialek, and M.J. Berry II. Synergy, redundancy, and independence in population codes. *Journal of Neuroscience*, 23(37):11539–53, 2003.

[30] Elad Schneidman, Susanne Still, Michael J. Berry, and William Bialek. Network information and connected correlations. *Phys. Rev. Lett.*, 91(23):238701–238705, Dec 2003.

[31] Giulio Tononi. An information integration theory of consciousness. *BMC Neuroscience*, 5(42), November 2004.

[32] Giulio Tononi. Consciousness as integrated information: a provisional manifesto. *Biological Bulletin*, 215(3):216–242, Dec 2008.

[33] Giulio Tononi. The integrated information theory of consciousness: An updated account. *Archives Italiennes de Biologie*, 150(2/3):290–326, 2012.

[34] Vinay Varadan, David M. Miller, and Dimitris Anastassiou. Computational inference of the molecular logic for synaptic connectivity in c. elegans. *Bioinformatics*, 22(14):e497–e506, 2006.

[35] Eric W. Weisstein. Antichain. `http://mathworld.wolfram.com/Antichain.html`, 2011.

[36] Paul L. Williams and Randall D. Beer. Nonnegative decomposition of multivariate information. *CoRR*, abs/1004.2515, 2010.

[37] Stefan Wolf and Jürg Wullschleger. Zero-error information and applications in cryptography. *Proc IEEE Information Theory Workshop*, 04:1–6, 2004.

[38] A. D. Wyner. The common information of two dependent random variables. *IEEE Transactions in Information Theory*, 21(2):163–179, March 1975.